FRIEDRICH NAUMANN FOUNDATION For Freedom.
Global Innovation Hub

TAMING THE DIGITAL REALM

Global Content Moderation Practices

Alexander Hohlfeld, Alphonce Shiundu, Ann Cathrin Riedel, Chung Ching Kwong, Dr. Gehan Gunatilleke, Mu-Huan Wang, Priscilla Ruiz Guillén

ANALYSIS

IMPRINT

PUBLISHER

Global Innovation Hub
Friedrich Naumann Foundation for Freedom
15F.-6, No. 171, Songde Road,
Xinyi District, Taipei City 110030
Taiwan

Web /freiheit.org/taiwan
Instagram /FNFGIHUB
Facebook /FNFGIHUB
LinkedIn /FNFGIHUB

AUTHORS

Alexander Hohlfeld, Alphonce Shiundu, Ann Cathrin Riedel, Chung Ching Kwong, Dr. Gehan Gunatilleke, Mu-Huan Wang, Priscilla Ruiz Guillén.

EDITOR

Global Innovation Hub of the Friedrich Naumann Foundation for Freedom

CONTACT

global.innovation@freiheit.org

AS OF

August 2023

NOTES ON USING THIS PUBLICATION

This publication is an information offer of the Friedrich Naumann Foundation for Freedom.

It is available free of charge and not intended for sale.It may not be used by parties or election workers for the purpose of election advertising during election campaigns (federal, state or local government elections, or European Parliament elections).

DISCLAIMER

The perspectives and opinions stated in this publication are those of the authors, and they do not necessarily reflect the view of Friedrich Naumann Foundation for Freedom.

TABLE OF CONTENTS

FOREWORD

Yu-Fen Lai
Program Officer "Digital Transformation"
Global Innovation Hub Taipei
Friedrich Naumann Foundation for Freedom

Dear Readers,

In this in-depth publication on global content moderation practices, we delve into the challenges and implications of regulating online platforms. Each case study sheds light on the unique sociopolitical context that shapes content moderation laws and their enforcement. By examining legislative attempts from Germany, the European Union (EU), the United Kingdom, Sri Lanka, Africa, Latin America, and Taiwan, we aim to provide insight into how governments are responding to the digital realm, where tons of information is created daily.

Like many other countries, in 2022, Taiwan was seeking to regulate online platforms and impose obligations on service providers regarding transparency requirements and content moderation. The Taiwanese government proposed a draft of the Digital Intermediary Services Act (DISA) but soon faced criticism from civil society groups, industry associations, and the general public. In response to these concerns, Taiwan swiftly suspended the process of launching the DISA, citing the lack of public consensus. This incident exemplifies the international influence of legislation adopted by leading democracies, as the DISA draft closely mirrored the Digital Service Act (DSA) of the EU. Yet, it also serves as a cautionary reminder that in jurisdictions where safeguards are not as sufficient, replication as such may compromise digital rights and freedom of expression online, if not become a tool abused by authorities. For instance, the German case study points out how authoritarian regimes, such as Russia and Singapore, invoke similar measures to the German Network Enforcement Act (Netzwerkdurchsetzungsgesetz, NetzDG) to oppress dissidents, albeit without incorporating the rule-of-law principle and checking mechanisms that exist in Germany.

Law enforcement for content moderation must consider the local sociopolitical and cultural context, as emphasized by the authors of the Latin America and Africa chapters in this publication. For example, in Kenya, both Meta and TikTok have outsourced the responsibility of checking online content to third-party services. However, these service providers failed to ensure that moderators understood the language, nor did they support them with sufficient mental health resources to cope with traumatizing content from their assigned tasks. Collaborative decision-making processes and compliance with international digital rights standards are also crucial aspects. Unfortunately, in Brazil, the civil society actors were not actively included in forming regulations combating fake news, leading to the question of whether such laws would lead to self-censorship and the future penalization of legitimate speech. Even worse, when content moderation outcomes contradict official narratives, some African governments deploy internet shutdowns and platform bans to punish social media platforms, as seen in Uganda and Nigeria.

Hate speech, disinformation, and harmful online content have long been evolving issues that threaten our democracy. However, governments need to strike a delicate balance between content moderation, holding digital platforms accountable, and upholding users' fundamental rights. For countries with poor track records in law enforcement, initiatives that empower stakeholders without granting additional regulatory power to the authorities could help create a better information ecosystem, as suggested in the case study of Sri Lanka.

In conclusion, this publication offers valuable insights into content moderation practices across different regions. By analyzing the successes, challenges, and potential pitfalls, we aim to contribute to the ongoing debates on creating effective and contextually legal frameworks that protect both freedom of expression and the well-being of users. We hope this publication can serve as a start for further discussions, policy-making, and international collaboration on the topic of content moderation.

DIGITAL SERVICES ACT: GRAPPLING WITH THE AMBIGUITIES OF DISINFORMATION

Alexander Hohlfeld

Introduction

The so-called Digital Services Act (DSA) of the European Union came into force in November 2022.[1] The DSA aims to "create a safer digital space where the fundamental rights of users are protected."[2] With massive developments in digital businesses and the digital public sphere, it was high time to replace the more than twenty-year-old e-Commerce Directive.[3] In the meantime, the preferred model of regulation for phenomena such as "hate speech" and "disinformation" was self-regulation.[4] However, these efforts, as exemplified by the Code of Conduct on Countering Illegal Hate Speech Online,[5] the Code of Practice on Disinformation,[6] and the Strengthened Code of Practice on Disinformation,[7] were evaluated as neither fully efficient nor satisfactory.[8] This can be explained by a number of public events, most notably the Cambridge Analytica scandal.[9] But also other incidents, such as the revelations regarding Facebook's role in Myanmar,[10] might have had an influence. Therefore, member states enacted their own regulations, with the German Network Enforcement Act (NetzDG)[11] standing out as a notable inspiration for the DSA. This article summarizes the key provisions laid down in the DSA on so-called "very large online platforms" (VLOPs), especially the obligation to carry out risk assessments, and highlights key challenges in the approach of the DSA in tackling disinformation.

Risk Assessments and Disinformation

The DSA proposal was published in 2020. Similar to the former e-Commerce Directive, the DSA includes a liability exception. This means that services are not held accountable for the content stored by users, as long as the services do not know of its existence. Instead, the approach of the DSA is to establish a framework of rules on transparency obligations, terms and conditions, complaint systems, and notice and action mechanisms, as well as on risk assessments and audits.

The framework established by the DSA can be regarded as a tiered system with different obligations for intermediary services, hosting services, and online platforms. The most stringent rules apply to VLOPs and "very large online search engines" (VLOSEs), i.e., online platforms with "average monthly active recipients of the service in the Union equal to or higher than 45 million."[12] On April 25, 2023, the European Commission designated 19 platforms, including Facebook, Twitter, YouTube, and TikTok, as VLOPs or VLOSEs.[13]

In terms of content moderation, the DSA differentiates between obligations on illegal content and obligations to evaluate and mitigate its influence on systemic risks, which do not necessarily need to include illegal content. According to the DSA, illegal content is not only that which "is not in compliance with Union law" but also content that violates the "law of any Member State which is in compliance with

Union law, irrespective of the precise subject matter or nature of that law."[14] This approach thereby challenges the intention of the DSA to harmonize the divergent national laws. The DSA introduces specific obligations to act against this content[15] as well as to put in place notice and action mechanisms,[16] enabling individuals to notify the platform about illegal content.

In addition to addressing illegal content, providers of VLOPs are required to carry out risk assessments[17] and mitigation measures.[18] These assessments involve evaluating whether the design or function of their service contributes to a range of broad categories referred to as systemic risks. These risks, including "the dissemination of illegal content through their services," "any actual or foreseeable negative effects" for the "exercise of fundamental rights," on "civic discourse," "electoral processes," and "public security," and in relation to "gender-based violence" and "the protection of public health and minors," as well as "serious negative consequences to the person's physical and mental well-being," need to be assessed.[19] However, the DSA lacks explicit guidelines on conducting these risk assessments, especially in terms of breaking down the broad and complex categories into specific risks that can be evaluated.[20]

The diffusion of disinformation is a highly debated issue in contemporary public discourse. Despite the omnipresence of the term, assessing and mitigating disinformation is much more complex and carries several risks. While the DSA does not offer disinformation an explicit definition, it mentions the term several times as a societal risk in the recitals, particularly concerning negative impacts on public health, public security, civil discourse, political participation, and equality, as well as the effects on minors.[21] It is therefore foreseeable that assessing whether platforms contribute to the spread of disinformation will be a key category of the risk assessments.

To determine whether content should be considered disinformation, both the truthfulness and the intention need to be assessed. However, in most cases, neither of these can be clearly legally evaluated, and even if the truthfulness could be evaluated unambiguously, wrong information too is covered by the fundamental right of freedom of expression.[22] As a result, implementing a complete ban or prohibition of disinformation would be neither possible nor desirable.[23] Nevertheless, the spread of disinformation can have serious consequences on fundamental rights, especially on the systemic risks mentioned. Therefore, the specific role of the platform, whether or not the design or function has a critical impact, needs to be evaluated. These risk assessments can then be requested by the European Commission or the Digital Service Coordinator.[24]

Conducting meaningful risk assessments requires an independent research landscape that observes the interdependencies between platform design, usage, and evolving risks. This is the only way to evaluate and critically assess the risk assessments carried out by the platforms. For this evaluation, transparency of the internal workings of platforms as well as comprehensive data access is needed. Prior to the existence of the DSA, researchers and civil society organizations had to rely on the goodwill of platforms to get necessary data access.[25] The DSA now steps in and establishes a framework of transparency and data access obligations. This framework enables independent research on the platform's inner workings and their contribution to societal, systemic risks.

When identified in risk assessments, platforms are obligated to implement mitigation measures that are reasonable, proportionate, effective, and tailored to the specific risks identified. At the same time, the impacts of these measures on fundamental rights need to be considered. The DSA mentions a broad variety of possible mitigation measures, such as adaption of the design of the services, terms and conditions, content moderation processes, and algorithmic systems.[27]

The selection of specific mitigation measures to tackle the spread of disinformation depends on the concrete assessment of the platform in its risk assessment. It would be conceivable, for example, to provide additional information on current issues or to modify parts of the recommender systems. Thereby, the impact on other fundamental rights, especially the right of freedom of expression, must be taken into account, making sure that the cure is not worse than the disease.

In addition to the general risk assessments and mitigation measures, the so-called Crisis Response Mechanism (CRM)[28] introduced by the DSA imposes further obligations on platforms. Through this mechanism, specific platforms can be obliged to carry out additional risk assessments and apply specific mitigation measures. This mechanism, incorporated late into the DSA's drafting process, has been considered one of the most critical points of the Act. Thirty-eight civil society organizations signed a joint statement and warned that "[t]he proposed mechanism is an overly broad empowerment of the European Commission to unilaterally declare an EU-wide state of emergency. It would enable far-reaching restrictions of freedom of expression and of the free access to and dissemination of information in the Union."[29] In response to these concerns raised by civil society, the final version of the DSA introduces several safeguards that prevent arbitrary use of the mechanism by the Commission. For example, the commission now requires the recommendation of the European Board for

Digital Services[30] to activate the mechanism.[31] However, criticism persists that the definition of what can be considered a crisis remains extremely broad, raising ongoing concerns that the use of the CRM could be more the rule than the exception.[32]

Conclusion

The DSA establishes a comprehensive framework of rules on transparency obligations, terms and conditions, complaint systems, and notice and action mechanisms, as well as on risk assessments and audits. The challenge of disinformation highlights the complexity of the obligations for VLOPs to conduct risk assessments and apply reasonable, proportionate, and effective mitigation measures. Risk assessments and mitigation measures need to be carried out cautiously, taking into account any unintended consequences.

An overreach of mitigation measures against disinformation may have enormous negative impacts on freedom of expression. It is therefore important to consider that more and more studies point out that the whole disinformation discourse can be characterized as a "moral panic,"[33] meaning that both the quantity and the influence of disinformation are overestimated in many Western societies, such as Germany, the United States, France, and the United Kingdom.[34] Moreover, the justification for policing disinformation is often used to monitor and silence criticism of government policies. This worrying trend is not limited to autocratic governments[35] but also occurs in established democracies.[36]

Instead of taking arbitrary content moderation measures, platforms should prioritize mitigation strategies that empower users. One common approach is to provide users with additional information on current issues. It is important to recognize that the problem of disinformation has existed throughout history. Tackling it through content moderation is neither possible nor desirable. Social media platforms would have to act as arbiters of truth, exerting an extraordinary impact on public discourse. To prevent potential abuses of power, it would be desirable for legislation to implement further safeguards on freedom of expression. These measures could limit the excessive authority of platforms as well as ensuring that political pressures aimed at combating and censoring opposing views will be rejected.

To ensure fundamental rights, it is crucial to establish mechanisms for monitoring by civil society and researchers, who both play a vital role in identifying the interdependencies among different risks and their corresponding mitigation strategies, along with potential political pressures that might undermine the protection of these rights.

References

1 European Union, Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act) (Text with EEA Relevance) (2022). http://data.europa.eu/eli/reg/2022/2065/oj/eng

2 European Commission, The Digital Services Act Package, Text, Shaping Europe's digital future. European Commission, June 2, 2020. https://ec.europa.eu/digital-single-market/en/digital-services-act-package

3 European Union, 'Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market (Directive on Electronic Commerce) (2000). http://data.europa.eu/eli/dir/2000/31/oj/eng

4 Jaursch, J. EU-Regeln Für Facebook & Co.: Was Der Digital Services Act Bringen Könnte, in *Digitaler Wandel Und Zivilgesellschaft: Positionen Und Perspektiven*, ed. D. Milovanovic, T. Staiger, & S. Embacher, Engagement Und Partizipation in Theorie Und Praxis (Frankfurt/M: Wochenschau Verlag, 2023), 112–14.

5 European Commission, Code of Conduct on Countering Illegal Hate Speech Online (2016). https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

6 European Commission, EU Code of Practice on Disinformation (2018). https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation

7 European Commission, Strengthened Code of Practice on Disinformation (2022). https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

8 Heldt, A. P. EU Digital Services Act: The White Hope of Intermediary Regulation, in *Digital Platform Regulation: Global Perspectives on Internet Governance*, ed. T. Flew & F. R. Martin, Palgrave Global Media Policy and Business (Cham: Springer International Publishing, 2022), 70. https://doi.org/10.1007/978-3-030-95220-4_4

9 Confessore, N. Cambridge Analytica and Facebook: The Scandal and the Fallout So Far, *The New York Times*, April 4, 2018. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

10 Amnesty International, Myanmar: Facebook's Systems Promoted Violence against Rohingya; Meta Owes Reparations, September 29, 2022. https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

11 BMJ, Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act, NetzDG) (2017). https://www.bmj.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html

12 DSA, Art. 33

13 European Commission, Digital Services Act: Commission Designates First Set of Very Large Online Platforms and Search Engines (Brussels, April 25, 2023). https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413

14 DSA, Art. 3 h)

15 DSA, Art. 9

16 DSA, Art. 16

17 DSA, Art. 34

18 DAS, Art. 35

19 DSA, Art. 34

20 Meßmer, A.-K., & Degeling, M. Auditing Recommender Systems (Stiftung Neue Verantwortung, January 23, 2023). https://www.stiftung-nv.de/de/publication/auditing-recommender-systems

21 DSA, Recitals: 2, 9, 69, 83, 84, 88, 95, 104

22 Appelman, N. et al., Truth, Intention and Harm: Conceptual Challenges for Disinformation-Targeted Governance, *Internet Policy Review* (blog), May 16, 2022. https://policyreview.info/articles/news/truth-intention-and-harm-conceptual-challenges-disinformation-targeted-governance/1668; Ó Fathaigh, R., Helberger, N., & Appelman, N. The Perils of Legally Defining Disinformation. *Internet Policy Review 10*, no. 4 (November 4, 2021). https://policyreview.info/articles/analysis/perils-legally-defining-disinformation

23 Jaursch, EU-Regeln Für Facebook & Co.: Was Der Digital Services Act Bringen Könnte.

24 DSA, Art. 34

25 Kayser-Bril, N. AlgorithmWatch Forced to Shut down Instagram Monitoring Project after Threats from Facebook, *AlgorithmWatch* (blog), August 13, 2021. https://algorithmwatch.org/en/instagram-research-shut-down-by-facebook/

26 DSA, Art. 40

27 DSA, Art. 35

28 DSA, Art. 36

29 EDRi, A New Crisis Response Mechanism for the DSA, *European Digital Rights (EDRi)*, April 12, 2022. https://edri.org/wp-content/uploads/2022/04/EDRi-statement-on-CRM.pdf

30 DSA, Art. 61, independent advisory group of [national] Digital Services Coordinators on the supervision of providers of intermediary services

31 EDRi, The DSA Should Pave the Way for Systemic Change, *European Digital Rights (EDRi)*, July 5, 2022. https://edri.org/our-work/the-dsa-should-pave-the-way-to-systemic-change/; Buijs, D., & Buri, I. The DSA's Crisis Approach: Crisis Response Mechanism and Crisis Protocols, *DSA Observatory* (blog), February 21, 2023. https://dsa-observatory.eu/2023/02/21/the-dsas-crisis-approach-crisis-response-mechanism-and-crisis-protocols/

32 Buijs and Buri, The DSA's Crisis Approach.

33 Carlson, M. Fake News as an Informational Moral Panic: The Symbolic Deviancy of Social Media during the 2016 US Presidential Election, *Information, Communication & Society 23*, no. 3 (February 23, 2020): 374–88. https://doi.org/10.1080/1369118X.2018.1505934

34 Altay, S., Berriche, M., & Acerbi, A. Misinformation on Misinformation: Conceptual and Methodological Challenges, *Social Media + Society 9*, no. 1 (January 1, 2023): 20563051221150412. https://doi.org/10.1177/20563051221150412; Jungherr, A., & Schroeder, R. Disinformation and the Structural Transformations of the Public Arena: Addressing the Actual Challenges to Democracy, *Social Media + Society 7*, no. 1 (January 1, 2021): 2056305121988928. https://doi.org/10.1177/2056305121988928; Acerbi, A., Altay, S., & Mercier, H. Research Note: Fighting Misinformation or Fighting for Information?, *Harvard Kennedy School Misinformation Review*, January 12, 2022. https://doi.org/10.37016/mr-2020-87

35 The Economist, Censorious Governments Are Abusing "Fake News" Laws, *The Economist*, 2021. https://www.economist.com/international/2021/02/13/censorious-governments-are-abusing-fake-news-laws

36 Big Brother Watch, Ministry of Truth (Big Brother Watch, 2023). https://bigbrotherwatch.org.uk/campaigns/ministry-of-truth/

The challenge of disinformation highlights the complexity of the obligations for VLOPs to conduct risk assessments and apply reasonable, proportionate, and effective mitigation measures. Risk assessments and mitigation measures need to be carried out cautiously, taking into account any unintended consequences.

Alexander Hohlfeld

CONTENT MODERATION AND COUNTERING DISINFORMATION IN AFRICA – THE TOUGH CHOICES

Alphonce Shiundu

In some African countries, social media has been blamed for fueling social unrest and enabling political instability. In other African countries, it has been credited with promoting democracy. The consensus is that social media is a double-edged sword that can build up democracies and tear down dictatorships on the one hand, while fomenting instability and entrenching digital authoritarianism on the other. Samidh Chakrabarti, when he served as Product Manager for Civic Engagement at Facebook, called it a "fundamental truth" that social media "amplifies human intent—both good and bad."[1] Indeed, some authoritarian governments have appropriated social media to shore up their stranglehold on their citizens, to crackdown on dissidents, and even to restrict freedom of expression.[2]

When examining efforts being made by African governments regarding the dangers associated with social media, the focus is often placed on the issue of hate speech. African governments tend to loosely define hate speech as content that is incendiary, threatening, inciting, and abusive, exploiting social divisions based on ethnicity, religion, and race. By definition, the content typically aims to mobilize hateful and violent action against specific groups of people.[3] In addition, there are legitimate concerns about the use of social media to spread false information regarding divisive topics and contested realities, thereby provoking social unrest.

Instances where social media platforms have been identified as the guilty enablers of sociopolitical unrest and civil conflict in Africa abound. Notably, Ethiopia and Sudan have experienced such challenges. In Ethiopia, social media platforms such as Facebook, Twitter, and WhatsApp were used by rebels and government allies to spread inciting content and hate speech to fuel the civil conflict which hit the country[4] from November 2020 to November 2022.[5] Similarly, in Sudan, scholars have documented instances where consistent publication of pro-state propaganda polluted the democratic space, manipulated public opinion, and exacerbated divisions.[6]

Back to Ethiopia: During the Tigray conflict from 2020–2022, the government itself executed a disinformation operation[7] on the aforementioned social media platforms to discredit the rivals in the conflict, malign pseudo-independent analysts and foreign organizations, suppress reports about human rights abuses and extrajudicial killings, and galvanize public support for a military operation to quell the rebellion.[8] In addition, in order to stifle debate over the controversies, the government shut down the internet, imposing a communications blackout in the restive parts of Ethiopia, citing security exigencies. It is noteworthy that government-sponsored disinformation tactics in Ethiopia predate the advent of social media, as similar strategies were used during the secession war with Eritrea in the 1990s, albeit through leaflets and radio broadcasts.[9]

In West African countries, such as Niger, Mali, and Burkina Faso, false information has had real-world harmful consequences, including fomenting xenophobic attacks, vandalism, and loss of public trust in the governments.[10] For instance, in early 2022, a video of vandalism shot in Angola was circulated with a claim that it showed Burkina Faso's embassy in Mali being vandalized because the country had supported sanctions against Mali from the regional economic bloc, i.e., the Economic Community of West African States (ECOWAS).[11] In South Africa, false information posted on social media platforms lit up xenophobic violence and heightened community tensions in 2019.[12] In Zimbabwe's 2018 elections, given the government's very tight media controls, social media became an avenue for the opposition and civil society to mobilize their supporters and to campaign.[13] However, videos of police brutality shared on social media raised political tensions and sparked violent clashes between protestors and security forces..

These examples underscore the critical role of social media in shaping the political landscape in Africa and reveal that most African authorities have an uneasy relationship with social media platforms, especially during election periods or during political crises when the very survival of these governments is at risk.[14] In the decade following the Arab Spring protests in 2011, several African governments, including those of Burundi, Ethiopia, Chad, Gambia, Uganda, Congo, Democratic Republic of Congo, Cameroon, and Gabon, shut down the internet during electioneering periods or political crises. Other governments, such as the Nigerian regime, banned social media platforms, specifically Twitter, from operating in the country.[15] In Kenya, during the 2022 elections, the government threatened to kick Facebook out of the country for failing to stop the publication of harmful content.[16] In response to the risks posed by disinformation, particularly those teetering towards hate speech, the governments of Madagascar, Tanzania, Kenya, and Ethiopia have implemented laws addressing these issues. However, the laws do not include provisions that mandate social media companies to remove offensive content within a specified period after being notified that the content has violated the community standards. The absence of such provisions may be due to the fact that it is difficult for African governments to push social media platforms to swiftly remove offensive content, given the platforms' limited resources allocated to the continent and the meagre government resources and technical capabilities to detect such content. In addition, fostering collaboration between governments and social media platforms requires mutual trust and effective systems. While stricter regulations may be considered, there is a risk of unintentionally harming legitimate speech and hindering innovation.

Challenges of Regulating Social Media Platforms in Africa

Some African countries, such as Kenya, have introduced laws to regulate social media platforms.[17] Still, these laws and regulations are poorly implemented, given the cross-border nature of social media platforms and the borderless internet. In addition, these laws targeting social media platforms and users were vaguely drafted that they often pose threats to media freedom and free expression and expose users to human rights abuses.

Another important point is the linguistic diversity of Africa. With between 1,000 and 2,000 languages spoken on the continent, most of which are oral,[18] there is hardly any economic incentive for social media companies to invest in human moderators and in developing algorithmic interventions to vet the content put out in all these languages. Most of the incendiary conversations in the digital public sphere of Africa happen in indigenous languages, beyond the radar of the algorithmic police and content moderators. Without addressing platform accountability, false content, disinformation, and even hate speech will continue to thrive online.

Moreover, the homogenizing portrayal of Africa as geopolitically uniform, even though there are about 55 countries on the continent, overlooks the important cultural differences that complicate the application of content moderation policies. What may be considered offensive in one country may turn out to be acceptable in another. Without local knowledge and nuanced understanding possessed by human moderators, social media platforms struggle to maintain credibility as authentic and unbiased forums for public discourse.

Furthermore, African governments are highly sensitive to political matters and often respond to the social dynamics surrounding politically volatile subjects. Consequently, when content moderation decisions by social media platforms clash with state-sponsored disinformation campaigns, as witnessed in Uganda and Nigeria, accusations of bias and censorship arise. In both cases, where internet shutdowns and platforms bans were implemented, social media platforms themselves faced punitive measures in response to political tensions.

In conclusion, African governments employ three main approaches when dealing with social media. First, they strictly regulate social media through laws or digital rules, some of which undermine free expression. Second, they shut down the internet or ban specific platforms from operating within the domestic jurisdiction. And third, they create a repressive offline environment—abductions and arbitrary arrests of digital influencers and opinion shapers who post alterna-

tive views contrary to the official narrative—consequently stifling public debate.

However, very few African governments have adequately explored the question of platform accountability, because this approach requires a range of strategies, including raising public awareness, collaborating with international organizations, providing capacity building and technical support, promoting regional cooperation, and offering funding. If these efforts are combined, African governments can better create a safer and more responsible digital environment while protecting users' rights.

Collaboration and Criticism: Initiatives From Social Media Platforms

Social media companies should be obligated to combat misinformation and disinformation, because their platforms have immense reach and influence. The unchecked spread of false information can lead to significant harm, including public confusion, erosion of trust, and potential real-world consequences. It is essential for these companies to take responsibility and mitigate the negative impact of information pollution on society.

Advocates within African civil society emphasize the importance of local content moderation teams that are familiar with local languages, dialects, and cultural contexts.[19] However, social media platforms may raise concerns about the impartiality of these moderators, particularly in relation to sensitive topics like elections, which often involve ethnic or tribal affiliations. Due to the influence of strong family bonds, it becomes necessary for countries to have institutions to supervise these moderators to maintain the integrity of the content moderation process. This step will enable social media platforms to move an inch closer to their altruistic goals of fostering a digital sphere that is free from harmful content and allows for genuine and honest public discussions.

In addition, social media users in Africa often encounter difficulties when trying to report inciting or inappropriate content. Therefore, making these processes more user-friendly and more easily accessible in local languages is needed. However, an important question arises as to who will fund the public initiatives aimed at educating users on how to report violations of community standards and platform guidelines. Even with secured funding, there remains the challenge of ensuring active user participation in flagging false or inciting content. Therefore, social media platforms need to adopt a comprehensive approach that addresses these aspects to work toward more effective and culturally sensitive content moderation processes in African contexts.

While some social media companies have made attempts to address the issue of content moderation, many have outsourced the responsibility to third-party services that prioritize the quantity of posts being moderated over the quality. This approach has led to concerns regarding the working conditions and treatment of content moderators in African countries. For example, in Kenya, a content moderator employed by Sama, the company contracted by Meta for content moderation, filed a lawsuit against Meta over poor wages and other poor workplace practices.[20] Moderators at Sama were asked to watch traumatizing and unsettling videos without receiving mental health support. The absence of such support ultimately affects the quality of moderation. In addition, a 2022 report published by the Mozilla Foundation revealed that TikTok moderators were assigned to moderate videos that they didn't fully understand, undermining their ability to effectively assess and moderate the content. The same group of moderators was also pressured to watch up to 1,000 videos per day, an excessive workload that compromised the moderation efforts and rendered them ineffective.[21]

On the other hand, there were also cases in which social media platforms partnered with African experts, local nongovernmental organizations, civil society, and even government agencies to gain insights into the local context and shape their content moderation policies accordingly. This collaborative approach aims to ensure the content moderation guidelines are more responsive and appropriate to the local context of African countries and the specific needs of the societies. For instance, Facebook has collaborated with African fact-checkers[22] through the Meta Third-Party Fact-Checking Program and provided the tools and the algorithmic capability to reduce the spread of harmful false information, hate speech, and other misleading and incendiary content on the platform. These partners, such as Africa Check[23] in Kenya, Nigeria, South Africa, and Senegal, contribute local context, cultural nuance, and political sensitivities in debunking false content. While other platforms such as Twitter and TikTok also worked with these fact-checkers during elections,[24] it appears that Meta currently maintains a sustainable partnership to counter disinformation in Africa.

It is also encouraging to see discussions taking place among some platforms, academics, users, governments, and civil society actors in Africa. In Kenya, for example, industry stakeholders, with the backing of the United Nations Educational, Scientific and Cultural Organization (UNESCO), launched a coalition in March 2023 for social media content moderation to help with advocacy regarding platform responsibility.[25] Prior to that, civil society organizations formed the Council for Responsible Social Media to foster dialogue with the platforms and hold them accountable.[26]

Conclusion

While platforms have made the appeals process transparent, there is still room for improvement in terms of transparency and responsiveness. In addition, there have been criticisms regarding outcomes of these partnerships. As the cooperation looks great on paper, the proposals made regarding solutions often disappear in the bureaucracy within the technology firms. Many critics believe that those firms who join the discussions either engage solely for public relations purposes or they lack the authority to push through the implementation of the proposals from these partnerships into technology solutions or algorithmic interventions. Bridging the gap between the proposals and outcomes generated from the partnership remains a significant challenge.

Finally, the prospect of artificial intelligence (AI) offers a ray of hope. Social media platforms could foster collaborations with stakeholders in Africa to explore the realm of machine learning and natural language processing technologies. This offers the opportunity to gain a more profound understanding of regional languages and dialects, thereby facilitating the adaptation of automated content moderation systems to effectively operate within local contexts.

These initiatives and partnerships reflect the recognition by governments, civil society, and social media platforms of the value of collaboration in strengthening the information ecosystem. Through collaboration, they can leverage their respective strengths and expertise to create a more robust and reliable information environment that benefits users and society. This understanding of the value of collaboration paves the way for continued cooperation and the pursuit of innovative solutions to foster a more trustworthy digital space.

References

1 Meta. (January 23, 2018). Hard Questions: What Effect Does Social Media Have on Democracy? https://about.fb.com/news/2018/01/effect-social-media-democracy

2 Bradshaw, S., & Howard, P. N. (2019). The global disinformation order: 2019 global inventory of organised social media manipulation. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1209&context=scholcom

3 Maina, H. (2011). The prohibition of incitement to hatred in Africa. Office of the High Commissioner for Human Rights. https://www.ohchr.org/Documents/Issues/Expression/ICCPR/Nairobi/HenryMaina.doc

4 Chekol, M. A., Moges, M. A., & Nigatu, B. A. (2023). Social media hate speech in the walk of Ethiopian political reform: analysis of hate speech prevalence, severity, and natures. *Information, Communication & Society*, *26*(1), 218–237. https://www.tandfonline.com/doi/full/10.1080/1369118X.2021.1942955

5 For more on the Tigray Conflict see: https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2022)739244

6 Bhatia, K. V., Elhussein, M., Kreimer, B., & Snapp, T. (2023). Protests, internet shutdowns, and disinformation in a transitioning state. *Media, Culture & Society*, 0(0). https://doi.org/10.1177/01634437231155568

7 AFP. (December 22, 2021). Ethiopia's warring sides locked in disinformation battle. *France 24*. https://www.france24.com/en/live-news/20211222-ethiopia-s-warring-sides-locked-in-disinformation-battle

8 Assefa, G. (February 2, 2022). Four ways the Ethiopian government manipulates the media. https://africanarguments.org/2022/02/four-ways-the-ethiopian-government-manipulates-the-media/

9 Pateman, R. (1995). *Intelligence operations in the Horn of Africa* (pp. 49–71). Palgrave Macmillan UK.

10 Yue, J., Bako, H., Hampton, K., & Smith, K. (July 2022). Conflict and the online space in the Sahel. Search for Common Ground. https://www.sfcg.org/wp-content/uploads/2022/07/Issue-Brief-Conflict-and-the-Online-Space-in-the-Sahel-July-2022.pdf

11 New Vision. (2022). Fake news flooding troubled Sahel region. https://www.newvision.co.ug/category/world/fake-news-flooding-troubled-sahel-region-127131

12 Chenzi, V. (2021). Fake news, social media and xenophobia in South Africa. *African Identities*, *19*(4), 502–521. https:doi.org/10.1080/14725843.2020.1804321

13 Mare, A., & Matsilele, T. (2020). Hybrid media system and the July 2018 elections in "post-Mugabe" Zimbabwe. *Social media and elections in Africa, Volume 1: Theoretical perspectives and election campaigns*, 147–176.

14 Masango, M., & Fourie, P. (2017, May 17). African governments versus social media: Why the uneasy relationship? Retrieved from https://theconversation.com/african-governments-versus-social-media-why-the-uneasy-relationship-73292

15 Anyim, W. O. (2021). Twitter ban in Nigeria: Implications on economy, freedom of Speech and information sharing. *Library Philosophy and Practice*, 0_1-13. Available from https://digitalcommons.unl.edu/libphilprac/5975/

16 Miriri, D. (2022, July 29). Kenya orders Meta's Facebook to tackle hate speech or face suspension. *Reuters*. Retrieved June 20, 2023, from https://www.reuters.com/world/africa/kenyas-cohesion-watchdog-gives-meta-7-days-comply-with-regulations-2022-07-29/

17 Communications Authority of Kenya and National Cohesion and Integration Commission. (2017, July). Guidelines on Prevention of Dissemination of Undesirable Bulk and Premium Rate Political Messages and Political Social Media Content Via Electronic Communications Networks. Retrieved from https://ca.go.ke/wp-content/uploads/2018/02/Guidelines-on-Prevention-of-Dissemination-of-Undesirable-Bulk-and-Premium-Rate-Political-Messages-and-Political-Social-Media-Content-Via-Electronic-Networks-1.pdf

18 Maho, J. (2004). How many languages are there in Africa, really? *Globalisation and African languages: Risks and benefits*, 279–296.

19 UNESCO. (May 2, 2023). Promoting digital rights and inclusion through addressing content moderation for sustainable internet for all. https://www.unesco.org/en/articles/promoting-digital-rights-and-inclusion-through-addressing-content-moderation-sustainable-internet

20 Vallace, C. (May 10, 2023). How social media is fueling unrest in Africa. *BBC News*. https://www.bbc.com/news/technology-64541944

21 Odanga, M. (May 10, 2023). From dance app to political mercenary: How disinformation on TikTok gaslights political tensions in Kenya. Mozilla Foundation. https://foundation.mozilla.org/en/campaigns/kenya-tiktok/

22 Meta. (May 17, 2023). A Map of Meta's Global Third-Party Fact-Checking Partners. https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking/partner-map

23 Africa Check's advisory to publishers on what to do if content on their page is flagged. https://africacheck.org/dont-delete-what-do-if-your-facebook-or-instagram-post-has-been-rated-false

24 Twitter. (August 3, 2022). The 2022 Kenyan general election is happening on Twitter. https://blog.twitter.com/en_us/topics/company/2022/the-2022-kenyan-general-election-is-happening-on-twitter

25 UNESCO. (March 14, 2023). Kenya launches a national coalition to fight against harmful content on digital platforms. https://en.unesco.org/news/kenya-launches-national-coalition-fight-against-harmful-content-digital-platforms

26 Council for Responsible Social Media. (March 8, 2023). Statement. https://accountablebigtech.com/index.php/statement-council/

African governments employ three main approaches when dealing with social media. First, they strictly regulate social media through laws or digital rules, some of which undermine free expression. Second, they shut down the internet or ban specific platforms from operating within the domestic jurisdiction. And third, they create a repressive offline environment—abductions and arbitrary arrests of digital influencers and opinion shapers who post alternative views contrary to the official narrative—consequently stifling public debate.

Alphonce Shiundu

NETZDG: CONTROVERSIAL YET PIONEERING WORK FROM GERMANY AGAINST HATE SPEECH

Ann Cathrin Riedel

Hate speech and online harassment, especially on social media platforms, have been a phenomenon visible since long before 2017 in Germany. Marginalized groups, such as women, people of color, Jews, and people with disabilities, have probably been aware of the pervasive presence of disturbing content and even (death) threats since they visited the World Wide Web for the first time. In the 2010s, the internet was still, for many people in Germany, as former Chancellor Angela Merkel called it, "Neuland" (new territory). However, the digital landscape began to change significantly in early autumn 2017, when social media played an influential role in the German federal government's election campaign.

The rise of the far-right party, the "Alternative für Deutschland" (AfD), in light of the 2017 elections marked a turning point. The AfD was elected to the Bundestag for the first time in 2017, making use of social media extensively for propaganda and campaigning. At the same time, the usage of social media platforms, especially Facebook, increased among the German population. As more and more politicians joined these platforms seeking direct engagement with their (potential) voters, politicians observed, if not encouraged, a surge in hate speech and threats predominantly by right-wing users—sent both via direct messages and posted openly as comments. The enormous refugee influx to Germany in 2015 inflamed the hate against politicians.

This article will explore the origin of the German Network Enforcement Act (Netzwerkdurchsetzungsgesetz), commonly known as the NetzDG, which was one of the pioneering legal attempts worldwide to address the issue of online hate speech. This text will provide a brief overview of the law and the surrounding discussions, as well as the results observed following its implementation. In addition, the article will examine subsequent improvements made to the act and shed light on how the European Digital Services Act (DSA) drew upon the lessons learned from the German NetzDG.

Overview of the NetzDG

Already in 2015, then Minister of Justice Heiko Maas (Social Democratic Party, SPD), agreed with social media platforms such as Facebook, Google, YouTube, and Twitter, along with civil society organizations, on implementing a task force assigned with providing recommendations for the sustainable and effective handling of hate speech on the internet. The agreement, "Together against Hate Messages,"[1] contained central elements of the future NetzDG. As part of the agreement, social media platforms committed to implementing user-friendly mechanisms to promptly review and remove potentially illegal content within 24 hours. Initially, the first draft of the NetzDG also included a provision addressing fake news, influenced by the significant role disinformation played in the US presidential election in 2016—a topic that was widely discussed in German media and politics. The provision, however, was later removed from the final version of the law.

Due to these societal circumstances and the elections in September 2017, the discussions surrounding the NetzDG gained momentum and became heated. While civil society organizations and groups such as the German Lawyer Association acknowledged the issue of hate speech on social media platforms, they disagreed on the content of the NetzDG and its hasty adoption before the parliamentary summer break. In the draft justification of the law, the minister of justice argued that the platforms' self-commitment did not yield satisfactory results. He reasoned this claim by pointing out the deletion rates of reported content on the platforms, which were at that time 90% for YouTube, 39% for Facebook, and only 1% for Twitter. Despite pushback from civil society, the federal government, consisting of the grand coalition of the SPD and Christian Democratic Union/Christian Social Union (CDU/CSU), passed the NetzDG in the final session of parliament before the summer break.[2]

It is noteworthy that the NetzDG did not changed the definition of legal or illegal speech in Germany. The law should only lead to the *enforcement* of criminal offenses on social media platforms. It imposes obligations on platforms to deal with user complaints, a principle also applicable to the DSA. Unlike the DSA, the NetzDG, however, exclusively affects social media platforms where users can share content with others and/or the public at large. Platforms for individual communication, such as messaging apps, gaming platforms, journalistic outlets, and online marketplaces, remained unaffected by the NetzDG. In addition, only platforms with a user base of at least two million are subject to the provisions of the NetzDG.

NetzDG in Practice

The main criticism against the NetzDG centered around the responsibility placed on social media platforms and their content moderators with no legal background, who are obligated to determine the legality of reported content and whether it should be deleted within 24 hours. This time frame applies to evidently illegal content, while less obvious content is given a seven-day time frame. Critics, including the then UN special rapporteur on freedom of opinion and expression, David Kaye,[3] expressed concerns that the right to freedom of expression is limited due to overblocking. According to the NetzDG, platforms must check and remove 22 offenses specified in the criminal code within the designated time limits. These offenses include both collective legal interests, such as those of the democratic constitutional state, and individual legal interests, like sexual self-determination, honor, personal life, and privacy.[4]

Platforms are not fined for failing to recognize and delete reported illegal content. Only an absent or insufficient governance system to handle user complaints is subjected to fines. The NetzDG did not take into account existing terms and conditions or community guidelines and how to deal with them. The DSA, on the contrary, takes these factors into account. Platforms affected by the NetzDG are obliged to publish a transparency report every six months, detailing the number and reasons for complaints, the duration of proceedings, and deletion rates. Another aspect that is subject to fines is the obligation for all social media platforms, regardless of their size, to designate a delivery agent (Zustellungsbevollmächtigten). This delivery agent acts as the main contact for authorities, fines, and civil court proceedings, as well as for requests for information from law enforcement agencies. This requirement of the NetzDG was and still is widely accepted, even among critics of the law, and has received consistent support.

The NetzDG came into effect on October 1, 2017, shortly after the German federal elections. The German Federal Office of Justice (BfJ), a subordinate authority to the Federal Ministry of Justice, expected an enormously high number of complaints and fine proceedings. By mid- 2020, the authority had initiated only 1,462 fine proceedings. Out of these, 1,353 were complaints directly submitted by users to the BfJ.[5] Approximately half of the proceedings were discontinued by the authority, either because the reported content was not illegal under the NetzDG and/or because no systemic failure by the platforms to handle user complaints was identified. To date, only one penalty notice has been issued under the NetzDG. This penalty was imposed on Facebook due to the complaint form for the NetzDG being positioned in a manner that made it difficult for users to find, thereby pressuring them to make complaints in accordance with community standards. Moderation decisions on these standards are not required to be included in transparency reports. Facebook was fined two million euros as a result of this incident.[6]

The Outcome of NetzDG: Ambiguous

It is hard to assess whether the NetzDG is a success due to a lack of comprehensive data for a profound empirical analysis.[7] Nevertheless, attempts to evaluate the law have come to completely opposite conclusions. A lot of illegal content remains virulent on social media platforms, even though platforms, especially the big ones (including Twitter before Elon Musk took over), make genuine efforts to address this issue. What can be said is that public pressure, amplified by the discussions surrounding the NetzDG, has made platforms more attentive to content moderation.

However, it is crucial to recognize that merely deleting illegal content, as determined by platforms, should not suffice in a constitutional state. The illegality must ultimately be determined by a court, and perpetrators should face consequences beyond the mere removal of content. For most criminal offenses, reporting to the police is necessary.

There is still much progress to be made in this area, as victims—especially women and marginalized people—have long complained about not being taken seriously at police stations and being advised to delete their own accounts or to take a break from social media. While there has been increased awareness within the police and justice system in recent years, online hate crimes are still not taken as seriously as they should be. In addition, the police often lack the appropriate knowledge and tools to effectively identify online perpetrators. But here too it is worth mentioning that the cooperation of platforms in providing data, despite the mandatory service of process, still requires improvement.

Furthermore, it is noteworthy that the expected "overblocking" resulting from the NetzDG has likely not occurred, although conclusive studies on this matter are lacking. One highly negative aspect stemming from the NetzDG is its international influence of the law, which is readily downplayed in the German debate, if not overlooked. Shortly after the implementation of the NetzDG in Germany, the Russian Duma adopted a similar law.[8] Similarly, several Asian and Latin American countries drew inspiration from the NetzDG, albeit without incorporating the rule-of-law mechanisms that exist in Germany. This international replication of the NetzDG raises concerns, particularly in jurisdictions where similar rule-of-law safeguards may be lacking.

Concluding Remarks: Lessons Learned from NetzDG and Path Forward

In the years following its implementation, several changes were made to the NetzDG, which cannot be discussed in detail here due to the length of this article. However, all of these amendments were met with fierce protests from civil society and business, as there were concerns that civil rights would be drastically curtailed. One amendment, for example, proposed that platforms directly forward content they deemed illegal, along with all of the user's associated data, to the Federal Criminal Police Office (BKA). This disproportionate measure and the definition of platforms as "deputy sheriffs" faced criticism not only from platform operators but also from net-political associations (to which the author of this article also belongs), the internet industry association eco, and civil society actors such as Reporters Without Borders,[9] to name just a few. Ultimately, platforms were required to report content to the BKA but were not obligated to pass the data of all flagged content to the BKA.

The DSA is set to largely, and probably entirely, replace the German NetzDG. The DSA has drawn valuable insights from the NetzDG, and its creators have carefully examined the pioneering German law. As a result, the DSA has not generated nearly as much controversy. Nevertheless, the business community and civil society have contributed to the debate surrounding its design with constructive and critical input. A study conducted by the Friedrich Naumann Foundation for Freedom from early 2022 also indicated that the DSA is likely to supersede the NetzDG[10]

The adoption of the DSA has been widely welcomed, given its significantly broader scope of application and regulatory impact. For example, unlike the NetzDG, the DSA covers all platforms and all forms of illegal content. Moreover, the DSA also extends its applicability to the gaming sector, an area often confronted with extensive hate speech. The DSA also places greater emphasis on user rights, including a right derived from Article 20 (4) to the restoration of content erroneously deleted by the platform and the enforcement of due diligence obligations through terms and conditions.

In conclusion, German lawmakers should bear three key considerations in mind. First, combating illegal content and violence in the digital space requires an ecosystem dedicated to addressing these issues. Relying solely on platforms is insufficient; police and judicial competencies are indispensable, as is the digitization of the judicial system for swift and competent action. Accelerating judicial processes is equally necessary. Second, criticism from the business community, civil society, and even the United Nations should be taken seriously. Rushing legislation, particularly when it involves potential restrictions on fundamental rights, should be avoided. Improved laws also enhance support for victims of digital violence. Finally, Germany often serves as a pioneer in legislation, as seen with the German Data Protection Act, which influenced European legislation, the General Data Protection Regulation. Germany sets an example on a global scale. While it is impossible to prevent nondemocratic governments from copying and abusing laws, it should give democracies pause for thought when such replication happens immediately. It is important to consider whether a law could be abused or whether sufficient safeguards are built in. If a UN special rapporteur feels inclined to interfere with German law, this could serve as an indication that there is room for further improvement.

References

1 Bundesministerium der Justiz. (December 15, 2015). *Gemeinsam gegen Hassbotschaften.* https://www.bmj.de/SharedDocs/Downloads/DE/News/Artikel/12152015_TaskForceErgebnispapier.pdf;jsessionid=CF49A654EEBB9EACCB59EAEA30CFB69A.1_cid334?__blob=publicationFile&v=2

2 *Entwurf eines Gesetzes zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken* (Netzwerkdurchsetzungsgesetz – NetzDG). (2017). Deutscher Bundestag 18. Wahlperiode. https://dserver.bundestag.de/btd/18/123/1812356.pdf

3 *Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, OL DEU 1/2017*. (2017). https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf

4 Peukert, A. (2022). Das Netzwerkdurchsetzungsgesetz: Entwicklung, Auswirkungen, Zukunft. In I. Spiecker gen. Döhmann, M. Westland, & R. Campos (Eds.), *Demokratie und Öffentlichkeit im 21. Jahrhundert – zur Macht des Digitalen* (pp. 229–248). Nomos. https://www.nomos-elibrary.de/10.5771/9783748932741-229.pdf

5 Ibid.

6 Wieduwilt, H. (July 2, 2019). *Millionen-Bußgeld für Facebook*. Frankfurter Allgemeine. https://www.faz.net/aktuell/wirtschaft/hass-im-netz-millionen-bussgeld-fuer-facebook-16265205.html

7 Rudl, T. (March 24, 2021). *Studie zeigt Schwächen bei Gesetz gegen Hassrede auf*. Netzpolitik.org. https://netzpolitik.org/2021/netzwerkdurchsetzungsgesetz-studie-zeigt-schwaechen-bei-gesetz-gegen-hassrede-auf/

8 Mchangama, J., & Fiss, J. (2019). *The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship.* Justitia. https://justitia-int.org/wp-content/uploads/2019/11/Analyse_The-Digital-Berlin-Wall-How-Germany-Accidentally-Created-a-Prototype-for-Global-Online-Censorship.pdf

9 Krempl, S. (February 13, 2020). *NetzDG-Reform: "Verdachtsdatenbank nie gekannter Dammbruch"*. Heise Online.https://www.heise.de/news/NetzDG-Reform-Verdachtsdatenbank-nie-gekannter-Dammbruch-4659517.html and https://www.reporter-ohne-grenzen.de/fileadmin/Redaktion/Dokumente/Stellungnahme_von_Reporter_ohne_Grenzen__RSF__zu_den_Gesetzesentwuerfen_zur_Bekaempfung_des_Rechtsextremismus_und_der_Hasskriminalitaet_und_zur_AEnderung_des_NetzDGs.pdf

10 Weiden, H. (2022). *Mehr Freiheit und Sicherheit im Netz Gutachten zum Entwurf des Digital Services Act*. Friedrich-Naumann-Stiftung für die Freiheit. https://shop.freiheit.org/#!/Publikation/1201

It is hard to assess whether the NetzDG is a success due to a lack of comprehensive data for a profound empirical analysis. Nevertheless, attempts to evaluate the law have come to completely opposite conclusions. A lot of illegal content remains virulent on social media platforms, even though platforms, especially the big ones (including Twitter before Elon Musk took over), make genuine efforts to address this issue. What can be said is that public pressure, amplified by the discussions surrounding the NetzDG, has made platforms more attentive to content moderation.

Ann Cathrin Riedel

ENCRYPTION IS EITHER PROTECTING EVERYONE OR BROKEN FOR EVERYONE

Chung Ching Kwong

What Is the Online Safety Bill and Why Is the Government Pushing for It?

The current world is facing one very urgent battle—identifying and removing harmful, hateful, and illegal online content. Over the past years, government officials in the United Kingdom (UK) have expressed concerns regarding online services' insufficient efforts to tackle illegal content, particularly on the issue of child sexual abuse material (CSAM). Members of the UK Parliament are taking steps toward making the country the "safest place in the world to be online while defending free expression."[1]

At the time of writing, the UK Parliament is moving forward with its Online Safety Bill (the Bill) in the House of Lords. The bill is a piece of legislation, wide in scope, that will govern all online service providers, commonly known as user-to-user services, including WhatsApp, Signal, Twitter, and TikTok, as well as search engines that operate in the UK. The aim of this bill is to protect users from being exposed to illegal content and minors from potentially harmful content.

Consensus Among Survivors: Tech Being Part of the Solution

According to a report published by Children Rights International Network on children protection and privacy, which included interviews with child sexual abuse survivors, two different perspectives are highlighted regarding the issue of stronger technological development to address online CSAM.[2] On the one hand, some survivors emphasize the need for stronger technology to pre-screen content prior to uploading or sharing.[3] They argue that this is the ultimate goal, as it prevents the content from being circulated and seen by others. On the other hand, some people with lived experience of abuse are staunch privacy advocates who find it offensive that survivors are being used to further a political surveillance agenda.[4] According to those people, current proposals to protect children online leave the door open for abuse of power and drive harmful activities further underground, making them more difficult to detect.

The report also reveals consensus across the spectrum of interviewees regarding the central role of technology in addressing online child sexual abuse. Those approaching the issue from a child protection perspective recognized that technology plays a direct role in facilitating abuse, enabling the spread of CSAM on a vastly higher scale than before. They stressed the urgent need to address the strong technological aspect and to develop technical solutions to combat this issue. However, scholars like Professor Andy Phippen, a professor of digital rights at Bournemouth University working on the issue from a children's rights perspective, cautioned against overstating the potential role of technology and warned against seeing technology as the sole solution to the problem. Given the rhythm of change in the digital world, some technological solutions are needed. Yet technology alone cannot be a silver bullet,[5] as some

believe this can lead to exaggerated claims about the capabilities of technical proposals without sufficient evidence. Moreover, there has been skepticism raised about the preventive potential of technology in addressing online child sexual abuse and exploitation,[6] noting that the necessary data and information originate from real-world contexts. Among those critical of technology, there exists a debate concerning the appropriate role and legitimate use of specific technologies.[7]

The discussion above outlines the core of the debate surrounding the Bill. At first glance, the Bill seems to be an ideal solution to the problem of harmful online content. It attempts to locate and remove all types of harmful content across the internet. However, the Bill resorts to a solution that is open to interception and could undermine personal privacy and security, which would put users, especially minors, at even more risk. By doing so, the Bill fails to strike a balance between content moderation and privacy.

Even though the Bill primarily targets social media companies, its definition of "content" encompasses anything that is "communicated publicly or privately." This broad definition covers almost all online activity. By this definition, the law not only targets social media platforms, as intended, but also every other online service accessible to users in the UK. The Bill would almost certainly cover every service that is available on the internet and create several problems.

Requiring Even More Information Than Now

The prevailing direction of privacy legislation is to limit the personal information required from consumers by apps and websites. However, the Bill deviates from this trend, as it mandates users to create accounts and confirm their age.[8] This approach facilitates cross-platform tracking, increasing its effectiveness beyond current levels.

Given the wide scope of the Bill, one noteworthy requirement is the mandatory age verification process on all websites, services, and applications that offer user-to-user content or communication within the UK. However, in order to carry out what is required by the proposed legislation, the traditional age verification approach that we used to use to block access to explicit content is not sufficient. The new age check mechanism involved may require direct verification using a passport, credit card, or other means, which often entails the collection of metadata or sensitive personal information, such as facial recognition or behavioral profiling.

In the UK, up to 88% of local companies have suffered data breaches in the last 12 months; there are 65,000 attempts to hack SMEs, around 4,500 of which are successful on a daily basis. Just 31% of UK organizations have done a digital risk assessment in the last 12 months.[9] In 2023, the IT Governance blog identified 277 million breaches in January and 29,582,356 breaches in February.[10] If more data are being collected and processed for age verification purposes of both adults and minors, is the UK ready to take actions to keep the data safe? When data breaches happen, there are no effective remedies to undo the damage being done. The need for proactive measures to prevent breaches to protect users, especially minors, are being overlooked here.

Too Broad a Scope and Blurred Lines Between Private and Public Content Leading to a Ban of E2EE

Another important provision within the Bill is Clause 110 that mandates websites and applications to proactively prevent harmful content from appearing on messaging services. This clause means that online service providers are to scan all user-generated content on a regular basis. In addition, the Bill also includes provisions that grant the Office of Communication (Ofcom), the communications regulator in the UK, the authority to access private messages on encrypted platforms. As a result, the use of end-to-end encryption will not be allowed, thus hindering our right to privacy and being unlikely to effectively protect minors and ordinary users. The inclusion of Clause 10, as well as the granting of surveillance powers to Ofcom, raises significant concerns. These provisions for a government backdoor create vulnerabilities that malicious actors could exploit, thereby putting users at more risk and eroding their right to privacy.

The broad application scope of the Bill goes beyond social media platforms: It would also cover instant-messaging services. Users' intimate texts with loved ones and dark humor memes shared among friendship groups are treated as the same category as things they publicly share on social media for everyone to see. Signals, WhatsApp, ProtonMail, and the secret chats of Telegram will all come under the purview of the Bill. This is blurring the line between private and public domains and would violate UK citizens' right to a private life. Such practices would lead to the universal scanning of all user-generated content at all times. It is not compatible with encryption or the citizen's fundamental right to privacy.

In a recent letter by prominent UK digital rights organizations,[11] serious concerns were raised about the Bill's impact on privacy and security in the country. In the letter, the organizations quoted warnings from leading cybersecurity experts that the ban of end-to-end encryption (E2EE) will pose "serious security and privacy risks for all society, while the assistance it can provide for law enforcement is at best problematic."[12] The Bill would give new powers to online in-

termediaries to use "accredited technologies" for conducting mass surveillance and scanning of all citizens on private messaging channels. Undermining E2EE protections could expose UK businesses and individuals to online vulnerabilities, including the very groups that the Bill aims to protect. It could also exacerbate the problem of child safety. Abused minors, for example, require private and secure channels to report what is happening to them. In addition, since the rights to privacy and freedom of expression are closely linked, these proposals could impede free speech, a crucial feature of open societies that distinguishes the UK from oppressors who employ coercion and repression to achieve their objectives.

Similar efforts in other jurisdictions that follow the same logic also encountered similar challenges. The belief that a backdoor or other workaround to read encrypted messages can be designed exclusively for targeting bad actors and only applied to benevolent purposes is unfounded. For instance, the U.S. Congress attempted to create backdoors to encryption with the EARN IT Act, and the EU proposed scanning private chats, which could potentially result in the mandatory scanning of every private message, photo, and video.[13] Additionally, government agencies attempted to pressure Apple to create software scanners on every device to constantly check for child abuse images and report back to authorities.

However, the reality is that no backdoor to encryption can exist without the risk of exploitation by bad actors such as cyber criminals, rogue employees, domestic abusers, or authoritarian governments, thereby compromising the security and privacy of individuals. For example, China has made similar requests to all online service providers within the country to carry out scans and censor all "problematic" content that goes beyond CSAM.[14] The Chinese government's requirements have resulted in restrictions on freedom of speech, where companies like Apple had to give up E2EE altogether within the country. Similarly, other services providing encryption face barriers to entering China's market due to government pressure. These instances represent a clear violation of privacy and freedom of speech. The Bill, which allows for mass scanning and surveillance, raises similar concerns.

While UK lawmakers claimed they do not intend to ban E2EE, the Bill nevertheless provides limited options available for encrypted services to comply with the regulation. The approved methods for compliance include removing or weakening encryption, installing client-side scanning, or ceasing service altogether. Such compromises would replicate the mass surveillance systems brought to light by Edward Snowden, thereby undermining UK citizens' right to privacy.[15]

Furthermore, concerns have been raised about the effectiveness of mass surveillance in preventing crime or terrorism. Adding to these concerns, the implementation of client-side scanning, as demonstrated by Google's disastrous experience, has raised serious questions about striking the right balance between privacy and addressing illegal content. In a 2021 paper titled "Bugs in our Pockets: The Risks of Client-Side Scanning," 14 computer science experts emphasized the doubts surrounding the efficacy and potential drawbacks of client-side scanning. There is little evidence to support that mass surveillance is effective in preventing crime or terrorism.[16]

Moreover, the Bill's potential to impede free speech and restrict the ability of journalists and whistleblowers to uncover wrongdoing is another worrisome concern. Given that there is a large diaspora community of Hongkongers as well as foreign dissidents residing in the UK due to the erosion of fundamental rights in their homeland, the limitations on E2EE imposed by the Bill would place them at significant risk.

Delegating to Online Service Providers Never Ends Well

Another controversial point is the Clause 65 included in the Bill, which requires online platforms like Facebook to enforce their terms of service under the threat of government sanctions, including criminal liability and jail time for executives. This has led to confusion, as it outsources the responsibility of defining harmful content to private companies and encourages self-censorship. The consequence: Companies might adopt an overly cautious approach by removing legitimate and protected speech from their platforms. This is similar to the FOSTA-SESTA bills in the United States,[17] which were meant to prevent sex trafficking but resulted in broad censorship around any content associated with "promoting or facilitating prostitution." The fallout from these bills led to platforms like Craigslist and Reddit shutting down their sections, and smaller websites ceased operations. The current language of the Online Safety Bill could have a similar chilling effect on free speech.

Policy Recommendations

This Bill in the UK highlights the need for a balanced approach to online regulations. It is vital to recognize that privacy and protection are not mutually exclusive, and both principles should be upheld, especially for minors, who are to be recognized as fully formed subjects of rights.[18]

The UK government should adopt a balanced approach to online regulations that upholds both privacy and protection principles, particularly for minors, who should be recog-

nized as fully formed subjects of rights. The UK government should refrain from blanket banning encryption and remove that from the bill as soon as possible, instead regulating its use in a manner consistent with children's rights, considering specific contexts and experiences, and respecting the principles of legality, necessity, and proportionality. In the process of developing a better Online Safety Bill, an inclusive approach would be used to establish a comprehensive and feasible child protection ecosystem, emphasizing prevention, education, appropriate funding, staff training, and multidisciplinary approaches to foster cooperation and address the diverse needs of those impacted by the legislation.

A general ban on encryption for children is not advisable and may actually make them more vulnerable to exploitation and abuse. Instead, regulating the use of encryption in a manner consistent with children's rights is a more appropriate approach. When intervening with encryption, it is important to consider the specific context and experiences of the children, including those from disadvantaged backgrounds and marginalized groups—those the Bill seeks to protect. For instance, when dealing with cyberbullying, policymakers would need to take into account the political landscape and the existing legal frameworks regarding online harassment. They would also need to consider the economic implications of implementing measures to combat cyberbullying, such as the cost of implementing monitoring systems or providing support services. Social and cultural factors would involve understanding the attitudes and behaviors surrounding cyberbullying within different communities or demographics. Any intervention must be provided for by law, be clear and precise, and be limited to achieving a legitimate policy goal in the least intrusive way possible.

In terms of technology-related legislation, a multi-stakeholder approach involving participation from various actors in the decision-making processes is essential. This includes involving government agencies, law enforcement, users, minors, survivors, and civil society in decision-making processes. Lawmakers and governments should also implement the model of open government, a concept that is rooted in the principles of transparency, accountability, and public participation in decision-making. Initially introduced in the United States in the 1950s, open government aims to move away from the traditional centralized and closed mode of governance toward a more democratic and participatory approach.[19] Various countries have adopted different structures and methods of implementation, such as releasing open information and organizing citizen deliberation activities. In recent years, the most well-known version of this structure is the Transparency, Participation, Public-Private Partnership proposed by the Obama administration in the United States, followed by the Transparency, Participation, Accountability, Inclusion version of the Open Government Partnership. By involving relevant stakeholders in decision-making, policymakers can make decisions that have better outcomes and build greater public trust.

Overall, policymakers need to understand that the impact of their decisions on encryption goes beyond their own jurisdiction. The digital world is interconnected, and regulations in one region are bound to have ripple effects globally or in other areas. Hence, policymakers must make a conscious effort to comprehend these connections, including by engaging with experts from different jurisdictions, especially those in marginalized communities.

References

1 Woodhouse, J., Conway, L., & Lipscombe, S. (2023). *Online Safety Bill: Commons stages. House of Commons Library,* UK Parliament. https://commonslibrary.parliament.uk/research-briefings/cbp-9579/

2 Child Rights International Network & defenddigitalme. (2023). *Privacy and Protection: A children's rights approach to encryption.* Child Rights International Network & defenddigitalme. https://home.crin.org/readlistenwatch/stories/privacy-and-protection

3 Ibid. p.55.

4 Ibid.

5 Phippen, A. (March 23, 2022). *Protecting children in the metaverse: it's easy to blame big tech, but we all have a role to play*. LSE. https://blogs.lse.ac.uk/parenting4digitalfuture/2022/03/23/metaverse/

6 Scott, M. (January 26, 2023). *Digital Bridge: Privacy vs. child protection — End of online advertising? — Governments own AI rulemaking*. Politico. https://www.politico.eu/newsletter/digital-bridge/privacy-vs-child-protection-end-of-online-advertising-governments-own-ai-policy/

7 Livingstone, S., Third, A., & Lansdown, G. (2020) Children's rights in the digital environment: A challenging domain for evidence-based policy. In L. Green, D. Holloway, K. Stevenson, T. Leaver, & L. Haddon (Eds.), *Routledge Companion to Digital Media and Children*. London: Routledge. https://eprints.lse.ac.uk/103006/1/Livingstone_childrens_rights_in_digital_environment_accepted.pdf

8 Under Article 11 Safety duties protecting children (3) A duty to operate a service using proportionate systems and processes designed to—
(a) prevent children of any age from encountering, by means of the service, primary priority content that is harmful to children (for example, by using age verification).
(b) protect children in age groups judged to be at risk of harm from other content that is harmful to children (or from a particular kind of such content) from encountering it by means of the service.

9 Databasix. (n.d.). *Statistics on Data Breaches in the UK, 2020*. Databasix. https://www.dbxuk.com/statistics/data-breach-statistics

10 Irwin, L. (March 1, 2023). *List of Data Breaches and Cyber Attacks in February 2023 – 29.5 Million Records Breached.* IT Governance. https://www.itgovernance.co.uk/blog/list-of-data-breaches-and-cyber-attacks-in-february-2023-29-5-million-records-breached

11 Polk, R. (November 24, 2022). *70 organizations, cyber security experts, and elected officials sign open letter expressing dangers of the UK's Online Safety Bill.* Global Encryption Coalition. https://www.globalencryption.org/2022/11/70-organizations-cyber-security-experts-and-elected-officials-sign-open-letter-expressing-dangers-of-the-uks-online-safety-bill/

12 Abelson, H., Anderson, R., Bellovin, S. M., Benaloh, J., Blaze, M., Callas, J., Diffie, W., Landau, S., Neumann, P. G., Rivest, R. L., Schiller, J. I., Schneier, B., Teague, V., Troncoso C. (2021). *Bugs in our Pockets: The Risks of Client-Side Scanning*. arXiv. https://doi.org/10.48550/arXiv.2110.07450

13 Europol. (n.d.). *Child Sexual Exploitation.* Europol. https://www.europol.europa.eu/crime-areas-and-statistics/crime-areas/child-sexual-exploitation

14 In particular, the Network Security Audit Regulations stipulate that participants in network security audits should protect the commercial secrets and intellectual property rights of operators of critical information infrastructures and providers of products and services, and should not use the relevant content for purposes other than auditing, such as external disclosure, without the consent of the information provider; the Regulations also prohibit restricting or discriminating against foreign products and services.

If a member of the censorship mechanism considers that an internet product or service affects or may affect national security, the Office of Network Security Censorship may, after reporting to the Central Committee for Network Security and Informatization for approval in accordance with the procedures, initiate a network security censorship in accordance with Article 15 of the Network Security Censorship Regulations.

15 Edward Snowden, a former NSA contractor, released classified documents in 2013, revealing extensive global surveillance programs conducted by the US government. His disclosures exposed the widespread collection of personal data and communications, sparking a global debate on privacy and government surveillance. Snowden's actions ignited controversy, as some regard him as a whistleblower who exposed government overreach, while others considered him a traitor for disclosing classified information.

16 Kirchner, L. (November 18, 2015). *What's the Evidence Mass Surveillance Works? Not Much.* Pro Publica. https://www.propublica.org/article/whats-the-evidence-mass-surveillance-works-not-much

17 Allow States and Victims to Fight Online Sex Trafficking Act of 2017, H.R.1865, 115th Cong. (2018). https://www.congress.gov/bill/115th-congress/house-bill/1865/text

18 CRIN and ddm interview with the Alana Institute, 22 September, 2022

19 Yu, H., & Robinson, D. G. (2012). *The New Ambiguity of 'Open Government'.* 59 UCLA Law Review, 59, 178. http://dx.doi.org/10.2139/ssrn.2012489

Overall, policymakers need to understand that the impact of their decisions on encryption goes beyond their own jurisdiction.
The digital world is interconnected, and regulations in one region are bound to have ripple effects globally or in other areas.
Hence, policymakers must make a conscious effort to comprehend these connections, including by engaging with experts from different jurisdictions, especially those in marginalized communities.

Chung Ching Kwong

THE REGULATION OF SOCIAL MEDIA PLATFORMS IN SRI LANKA

Dr. Gehan Gunatilleke*

*The author wishes to thank Devinda de Silva for his research assistance in compiling this paper.

Article 14(1)(a) of Sri Lanka's Constitution[1] guarantees the freedom of speech and expression, including publication, for all Sri Lankan citizens. The freedom of speech and expression via social media are, therefore, protected under Sri Lanka's Constitution. This freedom, however, is not absolute. Article 15(1) of the Constitution permits restrictions that are introduced by law and for specific purposes, such as protecting racial and religious harmony and parliamentary privilege and preventing contempt of court, defamation, or incitement to an offense. Therefore, under Sri Lanka's constitutional framework, speech and expression on social media can be subject to restrictions for specific public purposes.

There is currently no special law in Sri Lanka that sets out restrictions on social media. In the absence of such a purpose-built law, speech and expression on social media are regulated through a constellation of laws, policies, and institutions. This paper examines the legislative, policy, and institutional frameworks relevant to social media and analyzes their practical application. It also explores the advantages and disadvantages of introducing new specialized legislation to regulate social media in Sri Lanka.

The Legislative Framework

There are at least eight pieces of legislation in Sri Lanka that can be applied to harmful content on social media. Although these laws do not refer explicitly to social media, their scopes cover certain types of speech and expression online. Moreover, these laws only impose liability on users. This feature distinguishes Sri Lanka's legal system from other jurisdictions in Europe and Asia, where direct liability is imposed on internet intermediaries, including social media platforms.

Many of the relevant laws in Sri Lanka were enacted prior to the promulgation of the Sri Lankan Constitution in 1978. These laws have been retained despite any possible inconsistency with the fundamental rights chapter of the Constitution.[2] Additionally, the relevant laws that were enacted after 1978 were never successfully challenged before the Supreme Court of Sri Lanka, which exercises limited jurisdiction to review bills in terms of their consistency with the Constitution.[3] However, as discussed below, some of these laws have been flagged for their incompatibility with Sri Lanka's international human rights obligations.

Police Ordinance

The Police Ordinance[4] is one of Sri Lanka's earliest pieces of legislation to criminalize certain types of speech and expression. Section 98 of the Ordinance criminalizes spreading false reports to alarm the public and create panic. The scope of this offense easily covers disinformation shared on social media platforms. The penalty for such an offense can result in imprisonment for up to 12 months. However, as of time of writing, this provision is yet to be used to target harmful content on social media.

Penal Code

Several offenses found in the Penal Code cover speech and expression on social media. For example, section 291B[5] criminalizes writing or making visible representations that intentionally insult a religion or the religious beliefs of persons. This provision, in addition to the International Covenant on Civil and Political Rights (ICCPR) Act discussed below, was relied upon in the arrest of writer Shakthika Sathkumara, who was taken into custody in April 2019 for publishing a fictional short story on Facebook. The story was considered offensive to Buddhism and the Buddhist clergy, as it depicted sexual abuse in Buddhist temples.[6] Similarly, section 120 of the Penal Code[7] criminalizes "exciting or attempting to excite disaffection," which applies to speech and expression in general and, therefore, covers social media. Although criticism of the government is not criminalized under section 120, in practice, the provision has been used to target criticism of the government.[8] For example, in 2021, an assistant commissioner of the Land Settlement Department was arrested for sharing a Facebook post criticizing the government over deforestation.[9] The police relied on section 120 to make the arrest.

Antiquities Ordinance

Under the Antiquities Ordinance,[10] section 31 criminalizes offensive acts committed toward ancient monuments that are held sacred or in veneration. Although not classified as such, the offense in many ways resembles the offense of blasphemy, which may be generally defined as an act that is sacrilegious or insulting toward a divine being or sacred object.[11] This law can also be deployed to target harmful content on social media. In January 2021, it was reported that law enforcement authorities utilized this law, along with the ICCPR Act (discussed below), to arrest activist Sepal Amarasinghe for allegedly making disparaging comments on YouTube about the Sacred Tooth Relic in Kandy, an object many Buddhists venerate.[12]

Public Security Ordinance (PSO)

The PSO grants the president of Sri Lanka the authority to declare a state of emergency and issue emergency regulations. Although the PSO itself does not refer to social media, on several occasions, emergency regulations issued under it have prohibited certain types of speech and expression on social media. For instance, following the anti-Muslim violence that erupted in the Kandy district in February 2018, a state of emergency was declared, and emergency regulations were issued. The violence was triggered by a traffic incident that resulted in the death of a Sinhalese truck driver and led to two fatalities as well as damage to four mosques and over 400 Muslim-owned businesses and homes.[13] One of these emergency regulations, Regulation 15, criminalized the spreading of false rumors, statements, or images likely to cause public alarm, public disorder, or racial violence.[14] Notably, this was the first time an emergency regulation explicitly mentioned social media as a medium of communication. A similar set of emergency regulations was issued in April 2019 after the Easter Sunday Attacks, where an Islamist group called National Thowheed Jamaat launched simultaneous suicide bombings against three Christian places of worship and three hotels. The attacks claimed the lives of over 250 persons.[15] Regulation 15 of the 2019 regulations prohibited disinformation and incitement via social media.[16] The same provision was found once again in the emergency regulations issued in May and July 2023, respectively.

Prevention of Terrorism Act (PTA)

Section 2(1)(h) of the PTA[17] criminalizes speech and expression that incite acts of violence or religious, racial, or communal disharmony. This provision clearly covers content on social media, including speech and expression that incite violence or promote terrorism.

Section 2(1)(h) has been utilized to prosecute and punish journalists, such as J. S. Tissainayagam, for publications in print form. Tissainayagam accused the Sri Lankan military of committing war crimes and was convicted for inciting communal disharmony through his writing. Although this case did not involve social media, it indicates the potential for section 2(1)(h) to be applied to punish individuals who make similar statements on social media platforms. Over the years, the PTA has come under heavy criticism for being incompatible with Sri Lanka's international human rights obligations. Various United Nations treaty bodies, such as the Human Rights Committee, have pointed out the incompatibility of the PTA with Sri Lanka's commitments under the ICCPR, and have called for its repeal.[18]

Sri Lanka Telecommunications Act (SLTA)

The SLTA has been used on a number of occasions to restrict access to certain social media platforms. The Act grants the minister of technology the authority to issue directions to the Telecommunications Regulatory Commission of Sri Lanka (TRCSL) to limit access to social media platforms. Section 5(f) of the Act[19] mandates the TRCSL to take regulatory measures prescribed by the government in the interests of national security, public order, and the defense of the country. Additionally, sections 66 and 69 of the Act[20] authorize the minister to issue written directions to the TRCSL and, in the event of a public emergency, prohibit or control the transmission of messages via telecommunication.

The TRCSL has used its powers under the SLTA to impose broad restrictions on access to social media platforms. For example, in early 2018, the TRCSL blocked social media platforms, including Facebook and WhatsApp, for several days following anti-Muslim violence in the Kandy district.[21] The government justified this action as necessary to prevent the spread of hate speech and incitement on these platforms. Then, in April 2019, once again, access to certain social media platforms was restricted by the TRCSL following the Easter Sunday attacks.[22] The government claimed that the measure was necessary to prevent the spread of rumors and false information that could exacerbate tensions and lead to further violence. In April 2022, during public protests outside the president's residence, the TRCSL ordered internet service providers to restrict access to all social media platforms for several hours.[23] On each occasion, the TRCSL maintained that it received directions from the minister of technology, who is usually the president of Sri Lanka.

Computer Crimes Act

The Computer Crimes Act penalizes unauthorized access to computer systems, data theft, and other computer-related offenses. Section 6(1) of the Act[24] prohibits an individual from using a computer to perform any function that will endanger national security, the national economy, or public order. Conviction under this offense could result in imprisonment of up to five years.

This Act was frequently deployed during the COVID-19 pandemic to arrest internet users accused of spreading disinformation through social media. For example, in April 2020, a dance instructor was arrested under the Act for allegedly spreading a rumor on Facebook that President Gotabaya Rajapaksa had contracted COVID-19.[25]

International Covenant on Civil and Political Rights Act (ICCPR Act)

Section 3 of the ICCPR Act[26] prohibits the advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence. The Act has been used to target content on social media, even when the actual relevance of the content to the offense of "incitement" remained doubtful.

For instance, in 2020, Muslim activist Ramzy Razeek was detained under the Act for a Facebook post criticizing the government policy of mandating cremation for individuals who died from COVID-19. Razeek's arrest was due to his use of the term "jihad" (meaning "meritorious struggle" in Arabic[27]) when calling for peaceful resistance against the policy.[28] The ICCPR Act was also deployed in the arrest of writer Shakthika Sathkumara for his short story on Facebook. Sathkumara remained in custody for several months until it became clear that his actions could not be legitimately prosecuted under the Act. Similarly, in January 2023, activist Sepal Amarasinghe was arrested under the ICCPR Act for his comments about the Sacred Tooth Relic in Kandy. Amarasinghe was eventually released in February after issuing an apology to the Buddhist clergy.[29] Then, in May 2023, stand-up comedian and human rights activist Nathasha Edirisooriya was arrested under the Act for a stand-up routine posted on YouTube.[30] Edirisooriya is accused of offending Buddhism due to an innocuous reference to the Lord Buddha in one of her jokes.

The cases of Sathkumara, Amarasinghe, and Edirisooriya do not appear to involve any actual incitement against any community, including the Buddhist community. Instead, the persons concerned were accused of expressing views deemed offensive toward Buddhism, the Buddhist clergy, or an object sacred to Buddhists. This trend prompted the UN Human Rights Committee to call upon the government to refrain from prosecuting and imprisoning journalists, media workers, human rights defenders, and other civil society actors under the ICCPR Act "as a means of deterring or discouraging them from freely expressing their opinions."[31]

In contrast, in the fifteen years since its enactment, section 3 of the ICCPR Act has not led to the conviction of a single inciter of anti-minority violence in Sri Lanka. The instigators of anti-Muslim violence in Aluthgama in 2014,[32] Gintota in 2017,[33] Digana[34] and Ampara[35] in 2018, and Gampaha and Kurunegala in 2019[36] have yet to face any form of accountability. In the Gintota, Digana, and Ampara episodes, the violence was incited by militant actors via social media. However, in each case, no one was convicted under the ICCPR Act.

Policy Framework and Key Institutions

Apart from the legislative framework, Sri Lanka has developed policy and institutional frameworks relevant to freedom of speech and expression on social media. For example, the National Information and Cyber Security Strategy[37] outlines the state's approach to addressing cybersecurity issues and has implications for the regulation of social media platforms. "Thrust 2" (a term used to describe each area of intervention in the Strategy) on Legislation, Policies, and Standards contemplates the introduction of new laws to deal with cybercrimes, including those committed via social media. It also contains a commitment to enact a new Cyber Security Act, although the proposed Act is yet to be tabled in parliament. According to Sri Lanka's Ministry of Defence, once enacted, the Act is expected to have a notable impact on the regulation of social media through the newly established Digital Infrastructure Protection Agency[38]

The National Information and Cyber Security Strategy would be implemented by a range of existing institutions, including the Computer Crimes Investigation Division of the Police, the Information and Communication Technology Agency, and the Sri Lanka Computer Emergency Readiness Team, which serves under the Ministry of Technology.

Voluntary Code of Practice

The regulatory framework governing social media platforms in Sri Lanka has encountered two main criticisms from civil society actors. First, given the foregoing discussion, existing laws are seen as being applied selectively against dissenting voices, leading to concerns about the Sri Lankan government's abuse of these laws. Civil society has, therefore, resisted government plans to introduce additional laws. For example, the Sri Lankan government announced plans to introduce a new Online Safety Bill to regulate social media in 2023. According to news reports, the Bill is expected to be modeled on Singapore's Protection from Online Falsehoods and Manipulation Act.[39] Actors from civil society were strongly against this initiative, primarily due to a lack of trust in the government's intention, given its track record with existing laws.[40]

Second, civil society actors have been critical of social media service providers for their failure to effectively moderate and remove harmful content, including hate speech, disinformation, and incitement to violence. For instance, following the anti-Muslim violence in March 2018, which was fueled by social media platforms, 13 civil society organizations wrote an open letter to Mark Zuckerberg of Facebook[41] expressing disappointment in the company's slow response and lack of transparency in dealing with the issue. Therefore, despite concerns about laws passed by the government, Sri Lankan civil society actors recognize the need for some form of regulation of social media platforms.

Partly in response to this dilemma, a collaboration between a group within Sri Lankan civil society and the Asian Internet Coalition resulted in the drafting of a Code of Practice for Self-Regulation by Global Social Media and other Tech Companies in 2022.[42] This industry-wide code aims to establish voluntary standards for tech companies to uphold and is due to be launched later in 2023. According to its preamble, the Code aims to "enhance people's safety and contribute to reducing harmful content and behaviour online." The Code is not meant to replace any laws that govern social media content, but instead aims to make tech companies that become signatories accountable to the commitments that are relevant to their products or services. These commitments include: respect for freedom of expression and other fundamental human rights; the promotion of online safety; the protection of user privacy; ensuring that the platform's response to potentially harmful content meets the standards of necessity, proportionality, and reasonableness; and ensuring transparency. The compliance of signatories would be monitored by an independent monitoring mechanism.

Civil society support for the Code may be driven by the fact that it seeks to hold tech companies accountable without granting additional regulatory power to the government of Sri Lanka. Therefore, some experts considered the Code as a possible answer to the dilemma referenced above. However, others may remain skeptical about its effectiveness, as similar codes implemented in other countries, such as New Zealand, have faced criticism for allowing tech companies to deflect accountability.[43]

Conclusion and Policy Recommendations

Sri Lanka's existing legal, policy, and institutional frameworks for regulating social media platforms lack coherence and have been marred by selective enforcement as well as potential abuse. There is a significant gap in holding tech companies accountable for the content published on their social media platforms. However, attempts by the government to introduce new laws to fill this gap are resisted by civil society due to the authorities' poor track record in enforcing existing laws. In this context, the following policy recommendations may be considered.

First, the Sri Lankan government ought to develop a clear law enforcement policy that promotes restraint and ensures that laws related to social media content are enforced in good faith and with due regard for fundamental rights, including the freedom of speech and expression. For example, strict adherence to guidelines recommended by the Human Rights Commission of Sri Lanka could help to avoid selective and abusive enforcement of the ICCPR Act[44] and foster trust between civil society and the government in the regulation of social media platforms.

Second, the government should create space for (and even actively support) non-state initiatives, such as industry codes of practice, civil society-led monitoring of social media platforms, and media literacy programs that are designed to enhance online safety. While many of these initiatives may still be in their nascent stages (e.g., the above-mentioned Code of Practice) or have limited reach, they have the potential to contribute to the creation of safer online spaces.

Finally, once trust and space have been established, the government of Sri Lanka can embark on a more comprehensive legal reform project. This project should include

the repeal of obsolete and problematic legal provisions that are currently in use and their replacement with modern, human rights-compliant laws governing the use of social media in Sri Lanka. This step would require careful consideration, consultation with experts and civil society actors, and a commitment to upholding fundamental rights while addressing the challenges posed by social media platforms.

References

1 The Constitution of the Democratic Socialist Republic of Sri Lanka. https://www.parliament.lk/files/pdf/constitution.pdf [last accessed 1 June 2023].

2 Ibid. Article 16(1) of the Constitution provides: "All existing written law and unwritten law shall be valid and operative notwithstanding any inconsistency with the preceding provisions of this Chapter."

3 Ibid. See article 121 of the Constitution.

4 Police Ordinance, No. 16 of 1865. https://www.lawnet.gov.lk/police-4/ [last accessed 1 June 2023].

5 Penal Code Ordinance, No. 8 of 1883. https://www.lawnet.gov.lk/penal-code-consolidated-2/ [last accessed 1 June 2023].

6 Arrest of writer Sathkumara sparks debate on freedom of expression, *Daily Mirror*. (April 12, 2019). http://www.dailymirror.lk/news-features/Arrest-of-writer-Sathkumara-sparks-debate-on--freedom-of-expression/131-165392 [last accessed 12 May 2023].

7 Penal Code Ordinance, No. 8 of 1883. https://www.lawnet.gov.lk/penal-code-consolidated-2/ [last accessed 1 June 2023].

8 The explanation in section 120 of the Penal Code specifies that it is not an offense "to point out errors or defects in the government."

9 Asst. Land Commissioner arrested for sharing a post on Facebook, *Lanka Leader*. (May 22, 2021). https://english.theleader.lk/news/1418-asst-land-commissioner-arrested-for-sharing-a-post-on-facebook [last accessed 12 May 2023].

10 Antiquities Ordinance, No. 9 of 1940. https://www.lawnet.gov.lk/antiquities-3/ [last accessed 1 June 2023].

11 See 'Blasphemy' in the *Merriam Webster Dictionary* (2013).

12 Sri Lanka MPs set aside differences to break out pitchforks over YouTube comment, *Economy Next*. (January 5, 2023). https://economynext.com/sri-lanka-mps-set-aside-differences-to-break-out-pitchforks-over-youtube-comment-108420/ [last accessed 12 May 2023].

13 Borham M., Kandy communal violence: Main suspect arrested, *The Daily News*. (March 9, 2018). www.dailynews.lk/2018/03/09/local/145064/kandy-communal-violence-main-suspect-arrested [last accessed 1 June 2023].

14 Emergency (Miscellaneous Provisions and Powers) Regulations, No.1 of 2018. http://www.documents.gov.lk/files/egz/2018/3/2061-21_E.pdf [last accessed 1 June 2023].

15 Constable P. & Slater, J. Brothers of Sri Lanka bombing mastermind said to be killed in safe house battle; Catholics watch Mass on TV. *The Washington Post,* (April 28, 2019). https://www.washingtonpost.com/world/asia_pacific/brothers-of-sri-lanka-bombing-mastermind-said-to-be-dead-in-safe-house-battle-catholics-watch-mass-on-tv/2019/04/28/426d1e9c-6932-11e9-a698-2a8f808c9cfb_story.html [last accessed 1 June 2023].

16 Emergency (Miscellaneous Provisions and Powers) Regulations, No.1 of 2019. http://www.documents.gov.lk/files/egz/2019/4/2120-05_E.pdf [last accessed 1 June 2023].

17 Prevention of Terrorism Act, No. 48 of 1979. https://www.lawnet.gov.lk/prevention-of-terrorism-3/ [last accessed 1 June 2023].

18 Human Rights Committee. *Concluding observations on the sixth periodic report of Sri Lanka*. (April 26, 2023). CCPR/C/LKA/CO/6, para. 17.

19 The Sri Lanka Telecommunications Act, No. 25 of 1991. https://www.lawnet.gov.lk/sri-lanka-telecommunications-2/ [last accessed 1 June 2023].

20 Ibid.

21 Sri Lanka blocks social media as deadly violence continues, *The Guardian*. (March 7, 2018). https://www.theguardian.com/world/2018/mar/07/sri-lanka-blocks-social-media-as-deadly-violence-continues-buddhist-temple-anti-muslim-riots-kandy [last accessed 12 May 2023].

22 Sri Lanka blocks social media after Easter Sunday bombings. *NBC News. (April 22, 2019). https://www.nbcnews.com/tech/tech-news/sri-lanka-blocks-social-media-after-easter-sunday-bombings-n996886 [last accessed 12 May 2023].*

23 Sri Lanka protesters defy curfew after social media ban, *The Guardian. (April 3, 2022). https://www.theguardian.com/world/2022/apr/03/sri-lanka-protesters-defy-curfew-after-social-media-ban [last accessed 12 May 2023].*

24 Computer Crimes Act, No. 24 of 2007. https://www.lawnet.gov.lk/act-no-24-of-2007/ [last accessed 1 June 2023].

25 Directress of a dancing institute remanded for spreading false news about the President, *Newswire. (April 6, 2020). https://www.newswire.lk/2020/04/06/directress-of-a-dancing-institute-remanded-for-spreading-false-news-about-the-president/ [last accessed 12 May 2023].*

26 International Covenant on Civil and Political Rights Act, No. 56 of 2007. https://www.lawnet.gov.lk/international-covenant-on-civil-and-political-rights-iccpr-act-no-56-of-2007-5/ [last accessed 1 June 2023].

27 See https://www.britannica.com/topic/jihad [last accessed 30 June 2023].

28 Drop charges against Sri Lankan social media commentator. *Amnesty International USA*. https://act.amnestyusa.org/page/79791/action/1?locale=en-US [last accessed 12 May 2023].

29 YouTuber Sepal released after apology in court. *Daily News*. (February 22, 2023). https://www.dailynews.lk/2023/02/22/law-order/297939/youtuber-sepal-released-after-apology-court [last accessed 12 May 2023].

30 No right to laugh. *DailyFT*. (May 30, 2023). https://www.ft.lk/ft_view__editorial/No-right-to-laugh/58-748881 [last accessed 1 June 2023].

31 Human Rights Committee. *Concluding observations on the sixth periodic report of Sri Lanka*. (April 26, 2023). CCPR/C/LKA/CO/6, para. 41.

32 Haniffa, F., Amarasuriya, H., & Wijenayake, V. *Where Have All the Neighbours Gone? Aluthgama Riots and its Aftermath: A Fact-Finding Mission to Aluthgama, Dharga Town, Valipanna and Beruwela* (Colombo: Law & Society Trust, 2014), 1.

33 Bastians, D. Gintota and the shadows of extremism. *Daily FT*. (November 23, 2017). www.ft.lk/opinion/Gintota-and-the-shadows-of-extremism/14-643843 [last accessed 1 June 2023].

34 Borham, M. Kandy communal violence: Main suspect arrested. *The Daily News*, (March 9, 2018). www.dailynews.lk/2018/03/09/local/145064/kandy-communal-violence-main-suspect-arrested [last accessed 1 June 2023].

35 Ampara tense following attack on shop and mosque. *The Sunday Leader*. (February 27, 2018). https://web.archive.org/web/20180302041739/http://www.thesundayleader.lk/2018/02/27/ampara-tense-following-attack-on-shop-and-mosque/ [last accessed 1 June 2023].

36 Srinivasan, M. Mobs attack mosques, Muslim-owned shops and homes in Sri Lanka's Kurunegala District. *The Hindu* (May 14, 2019). https://www.thehindu.com/news/international/mobs-attack-mosques-muslim-owned-shops-and-homes-in-sri-lankas-kurunegala-district/article27119473.ece [last accessed 1 June 2023].

37 Ministry of Digital Infrastructure and Information Technology. *National Information and Cyber Security Strategy 2019–2023* (2018). https://cert.gov.lk/documents/NCSStrategy.pdf [last accessed 1 June 2023].

38 Govt. to bring new laws to combat emerging cyber crimes. *defence.lk*. https://www.defence.lk/index.php/Article/view_article/837 [last accessed 1 June 2023].

39 Sri Lanka to introduce new laws to regulate social media. *Adaderana* (January 6, 2023). http://www.adaderana.lk/news/87400/sri-lanka-to-introduce-new-laws-to-regulate-social-media [last accessed 12 May 2023].

40 A new law to regulate social media in Sri Lanka: A response. *Sri Lankan Brief*. (March 2, 2023). https://srilankabrief.org/a-new-law-to-regulate-social-media-in-sri-lanka-a-response/ [last accessed 1 June 2023].

41 Open letter to Facebook: Implement Your Own Community Standards. *Groundviews*. (April 10, 2018). https://groundviews.org/2018/04/10/open-letter-to-facebook-implement-your-own-community-standards/ [last accessed 1 June 2023].

42 SAFEWebLK: Code of Practice for Online Safety – First Draft finalized, *factum.lk*. (November 4, 2022). https://factum.lk/safe-web-lk/safeweblk-code-of-practice-for-online-safety-first-draft-finalized/ [last accessed 1 June 2023].

43 Daalder, M. Govt harbours concerns over Netsafe's online code. *newsroom*. (August 19, 2022). https://www.newsroom.co.nz/govt-harbours-concerns-over-netsafes-online-code [last accessed 1 June 2023].

44 Human Rights Commission of Sri Lanka. *Legal Analysis of the Scope of Section 3 of the ICCPR Act, No.56 of 2007 and Attendant Recommendations* (2019). https://www.hrcsl.lk/wp-content/uploads/2020/02/ICCPR-Act-s.-3-Guidelines-English.pdf [last accessed 1 June 2023].
SAFEWebLK: Code of Practice for Online Safety – First Draft finalized, *factum.lk*. (November 4, 2022). https://factum.lk/safe-web-lk/safeweblk-code-of-practice-for-online-safety-first-draft-finalized/ [last accessed 1 June 2023].

Sri Lanka's existing legal, policy, and institutional frameworks for regulating social media platforms lack coherence and have been marred by selective enforcement as well as potential abuse. There is a significant gap in holding tech companies accountable for the content published on their social media platforms. However, attempts by the government to introduce new laws to fill this gap are resisted by civil society due to the authorities' poor track record in enforcing existing laws.

Gehan Gunatilleke

INTRODUCTION OF THE DRAFT DIGITAL INTERMEDIARY SERVICES ACT IN TAIWAN

Mu-Huan Wang

Legal Status Quo and Origin of the DISA

At the end of June 2022, the National Communications Commission (NCC)—the Taiwanese communications regulatory authority—proposed a draft of the Digital Intermediary Services Act (DISA) as a fundamental regulatory regime for various digital intermediary service providers (DISPs[1]), and launched public consultation on it. However, less than two months later, the NCC suspended the initiative due to significant public opposition. This article aims to review the controversies sparked by the draft.

Taiwan has not established specific regulations for social media platforms and search engines. The current law mainly prohibits different types of illegal content on the internet and is scattered throughout various administrative effect laws. The most representative example is Article 46 of The Protection of Children and Youths Welfare and Rights Act. It established the administrative notice-removal model, whereby internet platform providers shall remove or disable access to internet content that is harmful to the physical and mental health of children and youth after being informed by local governments.[2]

On the other hand, since Taiwan is at the forefront of the defense against cognitive warfare in the Chinese-speaking environment, executive and legislative branches of the Taiwanese government implemented a series of amendments across various domains to prohibit the dissemination of rumors and disinformation. However, the new rules mainly focus on punishing those speakers rather than creating removal procedures for digital platforms; therefore, it cannot be regarded as legislation with integrity.

In this context, the NCC has been committed to initiating internet governance legislation for a long time. The Draft Electronic Communications Act[3] and the Draft Digital Communications Act[4] were proposed in the Legislative Yuan in 2016 and 2017, respectively. Both drafts were similar and characterized as basic laws that only provided general principles without any administrative control, and they were ultimately not enacted.

By 2020, Taiwan had experienced several cases of deepfake pornography and private sexual materials spread via the internet.[5] The Executive Yuan believed that regulatory strength was limited; therefore, it instructed the NCC to discard the light-touch approach of the former two drafts and introduce new laws to establish a common procedure for reporting, notifying, removing, or taking down all types of illegal content. It is the genesis of the DISA that mitigates the rapid dissemination and continuing harm of illegal information.

Introduction of the Draft and Reactions

The DISA Proposal

After exploring internet-related regimes around the world, the NCC decided to refer to the proposal on the Digital Services Act (DSA) put forth by the European Commission in December 2020[6] in developing a "Taiwanese DSA." Therefore, in many aspects, the DISA is extremely similar to the DSA proposal.

First, the DISA adopts the regulatory framework of cumulative obligation established by the DSA. Second, its definitions of regulated service providers are identical to the DSA, covering digital intermediary services (intermediary services in the DSA), including mere conduit services, caching services, and hosting services, as well as online platforms and designated online platforms (referred to as "very large online platforms" in the DSA). Notably, the DISA has adjusted the user threshold for these platforms to 2.3 million active users in reflecting the population of Taiwan.

Third, in terms of regulatory measures, the DISA retains the safe harbor provisions of the previous two drafts in 2016 and 2017. These provisions offer legal protections to regulated service providers. In addition, DISA imposes obligations that are almost the same as those proposed in the DSA. These obligations include the requirement for DISPs to disclose basic information, terms of use, and transparency reports. DISPs are also required to designate representatives and publish decisions regarding content moderation in publicly accessible databases. Most importantly, DISPs must comply with information restriction injunctions (IRIs) and temporary alert orders (TAOs) issued by relevant authorities. Moreover, providers of hosting services must establish notice and action mechanisms and inform affected parties about decisions related to content moderation. Online platforms, being at the core of the regulation, are additionally obligated to provide outlets for internal complaint handling and out-of-court dispute resolution. These online platforms must give priority to notices submitted by trusted flaggers, suspend or ban users repeatedly abusing their services, ensure they have knowledge of their business customers for online trading platforms, and disclose key information of online advertising. These measures aim to enhance transparency, accountability, and user protection within the digital ecosystem.

Nevertheless, the NCC seems to consider that the regulation for very large online platforms in the DSA proposal is too harsh. As a result, it deliberately moderates some of these obligations with DISA, marking the most significant difference between the two proposals. Designed online platforms are still required to comply with obligations such as annual self-risk assessment, adoption of risk-mitigation measures, and disclosure of key information of the recommended system. However, the obligation to conduct an annual external audit was modified to only be conducted upon order of the NCC. Furthermore, obligations related to online advertising reposition and making necessary data accessible were not adopted.

In addition, in order to abate concerns about oppressive enforcement after enacting the DISA, a dedicated organization was planned. This organization shall act as a bridge between the NCC and regulated service providers to develop codes of self-regulation and even establish legal orders and administrative rules.

Obviously, the regulatory scope of the DISA is very broad in order to achieve the original mission assigned by the Executive Yuan, the executive branch of the Taiwanese government. It covers not only online platforms such as social media and video sharing platforms, but also includes instant messaging applications being used for sharing private sexual images and disinformation, which constitute digital intermediary services. In addition, for reasons such as the prevention of the African swine fever virus, auction platforms allowing the sale of prohibited meat products are regulated under the DISA.[7] Hence, the ambitious proposal has attracted a lot of criticism and public debate.

Major Critical Opinions

In light of the fact that industrial self-regulation cannot effectively mitigate social risks from the distribution of illegal information, civil society affirmed the necessity of advocacy of platform accountability by the Taiwanese government. However, the DISA has still attracted the following criticisms.[8]

Information Restriction Injunctions

The most contentious provisions of the DISA are the information restriction injunctions. Articles 18 to 20 provide that, if deemed necessary, competent authorities may apply to the court for an IRI to a regulated service to remove or disable access to specific illegal information in order to prevent or mitigate harm to public interests. According to the NCC, the IRI procedures reference the provisions-related injunctions in the DSA proposal[9] and the UK's Online Safety Bill.[11] It follows the principle of prior judicial review[10] in order to protect freedom of speech as much as possible in individual cases. Though the DISA does not actually expand any new illegal content nor grant administrative authorities the power to take down illegal information unless other administrative effect laws allow it,[12] criticisms have emerged in the media, with some accusing the government of suppressing freedom of speech. These concerns and

allegations have sparked debate surrounding the balance between DISA and individual liberties.

Besides, at the same time, the Legislative Yuan, the parliament of Taiwan, was deliberating the "draft amendments to four laws to protect against sexual violence crimes," which was enacted in early 2023. It created the "expanded" administrative notice-removal model in the Sexual Assault Crime Prevention Act and the Child and Youth Sexual Exploitation Prevention Act, providing that "an internet platform provider, an internet application service provider, or an internet access service provider who learns of any matters suspected to be any [related] crime from any internet content protection agencies, competent authorities, police agencies, or other agencies, shall spontaneously restrict access to, or remove, webpage materials related to any such crimes." Some argue that the judicial procedures of IRI are not as quick and convenient as the administrative notice-removal model; thus, they claim that IRIs are redundant.

Temporary Alert Orders

Another major critical opinion is about temporary alert orders (TAO), which are also related to IRIs. Article 18 of the DISA provides that, after competent authorities apply to the court for an IRI and before the court issues it, they can order DISPs to temporarily display appropriate warning information attached to the rumors or disinformation concerned to make it easier for users to identify the controversy. The scope of a TAO is not defined explicitly. But, according to the legislative explanation, it should be a content-agnostic reminder, such as "This information is currently being reviewed by the court under the IRI procedure." The critics, however, regard TAOs as authorizing executive branches to judge rumors or disinformation before the court ruling, which can easily be misused to target political opponents.

Imbalance Between Large and Small Service Providers

As mentioned, the DISA covers a wide range of service providers and imposes numerous obligations, thus raising concerns about the feasibility of regulation for overseas service providers and the potential burden of compliance cost for domestic service providers.

With respect to overseas service providers, the practice community worried that it would lead to another case similar to the complete withdrawal of the paid service of Google Android Market from Taiwan in 2011, caused by the Taipei City Government's strong enforcement of the Consumer Protection Act on Google.[13] For domestic service providers, it is important to note that they are smaller in scale with fewer employees compared to big digital platforms; hence, they argue that they may struggle to bear the compliance cost of the DISA. For example, the Taiwanese longstanding academic bulletin board system known as PTT, despite its large number of users, implied that it may be forced to terminate services if the DISA is passed.

Post-DISA Legislative Actions

In light of the strong opposition from industry, news media, and the public, coupled with the sensitive timing of the 2022 9-in-1 local elections in Taiwan, the Executive Yuan promptly put a halt on the DISA initiative in mid-August. Thereafter, both the Legislative and Executive Yuan requested the NCC to strengthen communication with industry stakeholders and the public in reconsidering policy directions.

As a result of the setback in establishing a common mechanism for dealing with illegal information, several ministries have initiated their own decentralized legislation[14] in the administrative notice-removal model. For instance, in addition to "draft amendments to four laws to protect against sexual violence crimes," the Ministry of Health and Welfare as well as the Financial Supervisory Commission proposed an amendment to the Securities Investment Trust and Consulting Act, which was passed in May 2023.[15] Under the new rules, internet platform providers, internet application service providers, internet access service providers, and other online media operators who become aware of illegal advertisements shall remove or disable access to such ads.

Moreover, the Central Election Commission also amended the Civil Servants Election And Recall Act and the Presidential and Vice Presidential Election and Recall Act in May 2023 in order to prevent exploitation of deepfake technologies in interfering with elections.[16] According to the new rules, candidates have the right to request the police to conduct expert examinations of deepfake audio and visual content disseminated online. Subsequently, candidates can request internet platform providers or internet application service providers to remove or disable access to such material based on the examination's findings. These amendments were introduced to safeguard the integrity of elections in the face of emerging challenges posed by deepfake technology.

Challenges for Future Legislation

The discussion on norms of internet governance in Taiwan is still at an early stage, where longstanding practice has primarily been self-regulation or a light touch. Compared to the EU, which has implemented the e-Commerce Directive for over twenty years, the DISA can be seen as a leapfrogging legislative approach. The executive branches, therefore, could perform better in terms of responsibility for reasoning.

Meanwhile, Taiwanese users are accustomed to "free" internet space, lacking consensus on how to regulate related services. In particular, the quality of past drafts proposed by different ministries has varied, and there was a lack of policy communication with the public, leading to large-scale criticism from time to time. The NCC's poor reputation on broadcasting regulation thus makes the public more resistant to any proposal to limit freedom of speech. In addition, the timing of the proposal of the DISA was very close to the election period, hence amplifying misunderstanding, misinformation, and conspiracy theories about the provisions. The Executive Yuan's subsequent halting of the legislation in an attempt to immediately cut its losses ended up stifling further in-depth and rational debate in civil society.

Therefore, with complex regulation in the entire internet industry, a more appropriate and longer period of time should be devoted to investigating the industry and convincing the public. After all, digital transformation should not be driven only by an authority like the NCC; the overall administrative branches, judicial system, industry, and the public should all participate collectively.

Policy Recommendations

It is recommended that civil society acknowledges the necessity of a comprehensive legal framework for platform accountability in Taiwan with less criticism of obligations beyond IRIs and TAOs during public consultation on the DISA. While the initial legislative attempt to regulate digital intermediary services in Taiwan may not be perfect, it can still serve as valuable material for subsequent reviews and the launch of new initiatives by civil society stakeholders.

In terms of the overall legislative approach, given that the scale of the Taiwanese market is not as large as in other major economies, it is important to ensure regulatory alignment with these larger economies to reduce the risk of big digital platforms withdrawing from the Taiwanese market. Regarding classifying obligations, this article suggests adopting a cornerstone-and-addition legislative approach. This approach entails the establishment of a *lex generalis* (general law) regulation for all digital intermediary services and sectoral hard-law as *lex specialis* (special law) regulation for special types of platforms or illegal content. Particularly, for removal of or restricting access to illegal information, this approach allows for specialized and expedited administrative removal procedures that take into account sector-specific policy considerations, and it simultaneously ensures the safeguarding of freedom of speech by embracing the principle of prior judicial review as the common removal mechanism. By adopting this approach, Taiwan can strike a balance between providing a robust framework for platform accountability and allowing for flexibility and adaptability to evolving digital landscapes.

References

1 The draft can be found at https://www.ncc.gov.tw/chinese/files/22081/5542_47882_220811_1.pdf.

2 The Protection of Children and Youths Welfare and Rights Act (兒童及少年福利與權益保障法), Art. 46 (2003, amended 2011). https://law.moj.gov.tw/ENG/LawClass/LawSearchContent.aspx?pcode=D0050001&norge=46.

3 See generally 電子通訊傳播法草案 [Draft Electronic Communications Act], 立法院第九屆第一會期議案關係文書 [Agenda Related Document of 1st session of 9th Legislative Yuan] (2016). https://lis.ly.gov.tw/lgcgi/lgmeetimage?cfc6cfcececcc8ccc5cdc7c8d2cdc6c8.

4 See generally 數位通訊傳播法草案 [Draft Digital Communications Act], 立法院第九屆第四會期議案關係文書 [Agenda Related Document of 4th session of 9th Legislative Yuan] (2017). https://lis.ly.gov.tw/lygazettec/mtcdoc?PD090411:LCEWA01_090411_00028.

5 See Benack, E., *Taiwanese Youtuber Uses Deepfake Tech to Make Celebrity Porn*, RADIO TAIWAN INTERNATIONAL (October 20, 2021), https://en.rti.org.tw/news/view/id/2006342; Huang, T. T. *Taiwan President Vows Action against Deepfakes Amid Celebrity Porn Case*, TAIWAN NEWS (October 19, 2021), https://www.taiwannews.com.tw/en/news/4318972; Jiang, Y. T., Chen, C. Y., & Chen, H. J. , 臉被偷走之後：無法可管的數位性暴力?台灣Deepfake事件獨家調查 [*Stolen Faces: Unregulated Digital Violence? An Exclusive Investigation into Taiwanese Deepfake Incident*], 鏡新聞 [MIRROR MEDIA] (May 6, 2021). https://www.mirrormedia.mg/projects/deepfaketaiwan/.

6 See *Commission Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act),* COM (2020) 825 final (Dec. 15, 2020) [hereinafter DSA proposal].

7 See 行政院農業委員會動植物防疫檢疫局 [Bureau of Animal and Plant Health Inspection and Quarantine, Council of Agriculture, Executive Yuan], 防堵非洲豬瘟入侵 網路電商平臺不得販售違規肉製產品 [Preventing African Swine Fever: Online e-Commerce Platforms Prohibited from Selling Illegal Meat Products] (September 1, 2021); Feng, Z. W. *Chinese Pork Snacks Found in Taiwan Amid African Swine Fever Fears*, TAIWAN NEWS (April 8, 2021). https://www.taiwannews.com.tw/en/news/4171462; Lin, C. N. *Canned Pork from Vietnam Banned*, TAIPEI TIMES (December 26, 2019). https://www.taipeitimes.com/News/front/archives/2019/12/26/2003728201; Huang, T. T. *Taiwan Sounds Alarm on South Korean Products over Swine Flu Fear*, TAIWAN NEWS (Sep. 25. 2019). https://www.taiwannews.com.tw/en/news/3784012

8 For criticism from the industry, civic groups and academic experts, see NCC, **數位中介服務法草案專區** *[Special Webpage for DISA].* https://www.ncc.gov.tw/chinese/news_detail.aspx?site_content_sn=5542&cate=0&keyword=&is_history=0&pages=0&sn_f=47882 (last accessed May 1 2023) (enumerating YouTube links to videos of the three public explanatory meetings on the draft by the NCC in August 2022).

9 DSA proposal, Art. 41(3), 65.

10 Online Safety Bill, Bill 285 2021–22, Sec. 123–127 (March 17, 2022). https://publications.parliament.uk/pa/bills/cbill/58-02/0285/210285.pdf.

11 The concept of principle of prior judicial review refers to the legal design where statutes require a judicial order before commencement of law enforcement.

12 In other words, if there are provisions like the notice-removal model in administrative laws, these can be given priority for application as a *lex specialis* (special law). Otherwise, administrative authorities shall apply to the court for an IRI, uniformly referred to by the DISA as a *lex generalis* (general law), before requesting removal of illegal information.

13 For the event description and related case briefs, see Huang, S. C., & Chien, Y. Y. *Tax Challenges Resulting from Cross Border On-line Transactions*, 38 FIN. & ECON. L. REV. 145, 158–159 (2016).

14 The concept of decentralized legislation refers to the legislative strategy where there is no overall planning or uniform design, progressively formulating sector-specific laws as various subsectors evolve.

15 **證券投資信託及顧問法** [Securities Investment Trust and Consulting Act], Art. 70–1, 113–1 (2004, amended 2023), https://www.ly.gov.tw/Pages/ashx/LawRedirect.ashx?CODE=01652; Executive Yuan, https://www.ly.gov.tw/Pages/ashx/LawRedirect.ashx?CODE=01652; see also Executive Yuan, *Next-Generation Anti-Fraud Strategy Guidelines*, Version 1.5 (May 16, 2023). https://english.ey.gov.tw/News3/9E5540D592A5FECD/df8e5cf1-d2ca-4e5f-83b3-501aeaf1dcfb; Kao, S. C. *Meta, Google Aiding in Removing Fake Investment Ads*. TAIPEI TIMES (May 26, 2023).

16 **公職人員選舉罷免法** [Civil Servants Election and Recall Act], Art. 51-3, 110 (1980, amended 2023). https://www.ly.gov.tw/Pages/ashx/LawRedirect.ashx?CODE=01177; **總統副總統選舉罷免法** [Presidential and Vice Presidential Election and Recall Act], Art. 47-3, 96 (1995, amended 2023). https://www.ly.gov.tw/Pages/ashx/LawRedirect.ashx?CODE=04318; Chen, G. *Lawmakers Calling for Amendment to Election and Recall Acts*, PTS ENGLISH NEWS (November 22, 2021), https://news.pts.org.tw/article/555125; Hsieh, C. L., & Liu, T. H. *Amendments to Election and Recall Acts Sent to Cabinet*, TAIPEI TIMES (December 14, 2022). https://www.taipeitimes.com/News/taiwan/archives/2022/12/14/2003790699

Taiwan has not established specific regulations for social media platforms and search engines. The current law mainly prohibits different types of illegal content on the Internet and is scattered throughout various administrative effect laws.

Mu-Huan Wang

MISALIGNED EXPECTATIONS: LESSONS FROM THE DISA DRAFT CONTROVERSY IN TAIWAN

Anonymous

The draft bill of the Digital Intermediary Service Act (hereinafter the DISA Draft)[1] was a failed rule-making attempt by the Taiwanese government (hereinafter the Government) to hold internet intermediaries, especially internet platforms, accountable. While the Government failed to further finalize the DISA Draft into a formal bill and submit it to the Legislature, the DISA Draft is still worthy of further discussion on its regulatory scope, its immediate impact, and its reverberations in society and the circle of policymakers: The DISA Draft provides insight into how the Government seeks to react to the current online environment, which it considers to be less safe, trustworthy, and reliable.

To achieve these goals, the DISA Draft incorporated provisions of the Digital Service Act of the European Union (hereinafter EU-DSA) to impose obligations, such as the general legal compliance requirements and the safe harbor provisions for the takedown of illegal contents, on all internet intermediaries.[2] Additionally, specific requirements were proposed for internet platforms, such as due process for user complaints and dispute handing for internet platforms,[3] requirements for addressing reports on illegal contents from users and trusted flaggers,[4] and transparency requirements for all platforms, some in particular for very large online platforms.[5] However, the most innovative yet controversial aspect of the DISA Draft, which ultimately led to its failure, was its mechanism for addressing disinformation. The DISA Draft proposed granting administrative agencies the authority to (i) obtain a court order to restrict access for certain illegal contents on all intermediaries,[6] or (ii) issue a temporary administrative order that would require platforms to display warning messages for content that the Government deemed to be rumors or disinformation.[7] These regulatory designs aimed to establish a framework defining basic principles and standardized procedures for intermediaries' interactions with users and the Government, but some of them did not receive recognition within Taiwanese society.

Prior Legislative Attempts Before the DISA Draft

Before all the controversies surrounding the DISA Draft emerged in 2022, the Government introduced the Digital Communications Act bill (hereinafter the DCA Bill) in 2017 to govern internet platforms. Initially, the submitted version of the DCA Bill[8] did not address the issue of disinformation, as it was not considered a priority at that time.[9] However, as disinformation gained prominence in Taiwanese society from the latter half of 2018, the Government proposed a revision of the DCA Bill to empower government agencies to require platforms to remove disinformation content that violated existing laws within a 24-hour time frame. This revision aimed to address the perceived inadequacy of the original bill in governing these platforms and the pressing issue of the infodemic.[10] However, the revision faced strong criticism from private sectors and human rights advocacy groups, leading the Government to withhold the revision.[11] Consequently, the DCA Bill expired in January 2020 as the legislative term ended.

Nevertheless, the Government remains committed to addressing the issue of illegal and harmful online content. It recognizes that relying solely on *ex-post* punishments is insufficient to halt the spread of disinformation from the outset.[12] Despite internet platforms launching self-regulatory campaigns against disinformation,[13] doubts have been raised by media scholars,[14] the general public,[15] and various stakeholders regarding the actual effectiveness of such efforts. Increasing instances of scams, questionable decisions regarding user complaints, and opaque content moderation practices have further fueled resentment toward internet platforms.[16] Moreover, as the EU-DSA draft was proposed and has been actively discussed worldwide since December 2020, it became a reference point for policy discussions in Taiwan. Its framework for governing internet platforms serves as potential justification for future regulatory designs in the country.[17]

Stakeholder Reactions and the Aftermath

When the Government released the DISA Draft, it received mixed responses. Initial concerns raised by experts, civil society groups, and opposition parties focused on the enforcement infeasibility and its potential impacts on freedom of speech. However, none of these concerns directly rejected the idea of increased platform regulations.[18] The peak of the concern regarding speech control was reached during public hearings held by the National Communications Commission (NCC) in mid-August, 2022. During these hearings, both international and local internet platforms strongly criticized the DISA Draft for incentivizing over-censorship, placing burdensome requirements on platforms operated by volunteers (including smaller local online forums and community-based collaboration platforms like Wikipedia), potentially undermining communication secrecy, and creating conflicts with existing regulations.[19] Their critiques, along with statements from industry associations,[20] fueled public discontent and raised apprehension over potential government intervention in political speech. In response to these concerns, the Government swiftly announced an indefinite halt to the rule-making process, citing the lack of public consensus.[21]

Following the halt of the DISA Draft rule-making process, government agencies initially refrained from explicitly mandating content takedowns. Instead, they sought voluntary assistance from platforms to moderate gravely harmful content that could exacerbate high-profile criminal activities or interfere with the integrity of the upcoming local elections in late 2022.[22] Concurrently, the Government adopted a "divide and conquer" approach alongside the DISA Draft, introducing specific legislations to address harmful content in different contexts, such as requirements for platform takedowns of child sexual abuse material (CSAM), nonconsensual pornography, investment fraud, and election-related deepfake audiovisuals.[23] These proposals received swift passage in the Legislature[24] without major opposition from internet platforms, due to relatively high public support for measures that ensure removal of such harmful content.[25] The possibility of the Government revamping and reintroducing the DISA Draft remains uncertain. Nonetheless, some politicians and online influencers argue that these legislations would have been unnecessary had the DISA Draft already been enacted,[26] suggesting that the failure of the DISA Draft was not only unwarranted but also the result of coordinated disinformation campaigns.

In contrast, civil society groups and scholars have adopted a more cautious stance when commenting on the DISA Draft following its controversies. While recognizing concerns about potential threats to freedom of speech resulting from over-censorship, some of these groups and scholars appreciate the requirements for increased transparency and due process in content moderation and handling user complaints, as well as the formalization of the court's role in content takedown procedures.[27] Some even argue that social media platforms, as online gatekeepers, should adopt a more proactive approach in addressing harmful content, and that relying solely on self-regulation may not be entirely feasible.[28] It is worth noting that these opinions were primarily expressed a few months after the controversies emerged. Overall, discussions from these sectors were lukewarm, if not timid, compared to the turmoil within society.[29] Their attitudes may be best captured by the following remarks made by media scholar and former NCC commissioner Chi-shen Ho:[30] "The DISA Draft, inspired by the latest legislation in the European Union, opens a window for policy discussions within Taiwan and should be taken seriously by all sectors. Unfortunately, the authorities did not gradually lead discussions and planning with industry stakeholders, and there was a lack of sufficient demonstration of tentative consensus and dialogue on various key issues. This led to a lack of trust in the government's legislative intentions. […] The formulation of relevant strategies should still involve patient and extensive communication with internet service providers and users."

Reflections and Policy Recommendations

The failed rule-making process of the DISA Draft did initiate the first formal discussion in Taiwanese society regarding the need to hold platforms accountable and has become a (potentially negative) reference for future policy development. The controversy clearly demonstrates the limitations of a traditional top-down approach to rule-making, where the government determines the agenda without sufficient consensus from various stakeholders and the general public, thus eroding the mutual trust. However, it remains

unclear how multi-stakeholderism, which has long been endorsed by the internet governance community, could be effectively implemented in the policy development process in Taiwan. Regardless of the specific process in the future, the author would like to remind policymakers that not all goals aimed at addressing internet platform issues can align without conflict in practice. Misalignments may be more common than expected, especially in the context of unavoidable automatic content moderation for large-scale online activities.[31]

For future policy makings, the author would like to suggest the following:

1. The DISA controversy in Taiwan clearly demonstrates that it may not be feasible to address multiple policy goals in a single legislation package like the EU-DSA. These goals may misalign, and depending on the context, hinder the progress of each other.
2. To achieve a proper balance of policy goals, it is necessary to engage in multiple rounds of discussions involving various stakeholders. The expertise and experience of internet platforms and integrity workers within these platforms should be considered.[32] This inclusive approach enables a comprehensive assessment of the feasibility of balancing different policy goals within a specific proposal.
3. For countries that lack bargaining powers when dealing with very large internet platforms in their smaller markets, it may not always be practically feasible for the authority to take the lead in formulating regulations. However, these countries can benefit from the regulations established by leaders such as the EU or other markets with greater bargaining power.[33] They can gradually develop their own rules that can be harmonized with those of the leading markets. This "hitchhiking" approach ensures that large internet platforms comply with similar requirements, as they should adhere to a uniform policy across all markets. To make this approach realistic, collaboration between leaders in regulations and these late-moving countries is necessary. This enables the sharing of experiences and addressing of concerns within the policy development process.

References

1 For the full text of the DISA Draft, see the archive by the National Communications Commission (NCC), the regulator of internet, telecom, and broadcast at that time: NCC. (June 29, 2022). *Shuwei zhongjie fuwu fa cao'an zongshuoming [General Description of the Digital Intermediary Service Act Draft].* https://www.ncc.gov.tw/chinese/files/22062/5532_220629_1.pdf

2 See Article 5 to Article 16 of the DISA Draft.

3 See Article 22 to Article 25, and Article 27 of the DISA Draft.

4 See Article 26 of the DISA Draft.

5 For transparency requirements for all platforms, see Article 28 to Article 21 of the DISA Draft; for requirements for very large online platform, see Article 32 to Article 33

6 See Article 18, paragraph I–VII, Article 19, and Article 20 of the DISA Draft. However, it shall be noted that instant messaging services are included in the definition of intermediaries, which is not exactly identical to the design of EU-DSA (which exclude interpersonal communications like instant messaging or email)—see Article 2, subparagraph (2), item 1, and relevant legislative memorandum (Article 2) paragraph III, subparagraph (1) of the DISA Draft. Such requirement, when applied to instant messaging services or other interpersonal communications, may incur communication interception, surveillance, and censorship that overlap with the communication wiretapping laws in Taiwan. Also, this provision does not require the court to consider the technical feasibility of access restriction; therefore, it may be non-compliable for certain services, e.g., those utilizing end-to-end encryptions (E2EE). Non-compliance of the court order, though it might be incurred from technical infeasibilities of the subjected services, could result in access restrictions/denials to the services as penalties. See Article 54 and Article 55 of the Draft.

7 See Article 18, paragraphs VIII–XI. Also, the author believes that the inspiration of this measure is worthy of further discussion. While the relevant legislative memorandum (Article 18, paragraph VIII) did cite the revised Directive (EU) 2018/1808 as the reference, such a measure was not present in the cited Directive. As far as the author is aware, the most similar measures are the "correction direction" in the Protection from Online Falsehoods and Manipulation Act 2019 (POFMA) and the Foreign Interference (Countermeasures) Act 2021 (FICA) of Singapore. The correction directions could be solely issued upon the discretions of government agencies and have been criticized for being biased

against political opposition to the ruling administration.

8 For the full text of the DCA Bill, see the archive of NCC: NCC. (April 18, 2017). *Shuwei tongxun chuanbo fa cao'an zongshuoming [General Description of the Digital Communications Act Draft]*. NCC. https://www.ncc.gov.tw/chinese/files/17041/3861_37260_170418_1.pdf

9 However, the DCA Bill did grant the users, digital communication service providers (including internet platforms), and affected third parties the right to seek injunctions in court in cases where disputes arose regarding the usage of the provided service (which may potentially include disinformation), while injunctions were necessary to prevent major harm, immediate dangers, or other similar situations. See Article 19 of the DCA bill.

10 Kao, S. S. (2021). *Taiwan's Response to Disinformation A Model for Coordination to Counter a Complicated Threat* (NBR Special Report no. 93). The National Bureau of Asian Research. https://www.nbr.org/wp-content/uploads/pdfs/publications/sr93_taiwan_sep2021.pdf

11 Ibid.

12 See ibid, 12.

13 See the news release by the Taipei Computer Association, a trade association of which Meta, Google, LINE, and Yahoo Taiwan are all members: Taipei Computer Association. (June 21, 2019). *Zilyuxianxing benhui yu sida pingtai yezhe xishou fangzhi bushixunxi [Self-regulation is the first step. This Association and the four major platform operators work together to prevent misinformation]*. Taipei Computer Association. https://www.tca.org.tw/tca_news1.php?n=1411 For the Code of Conduct of this self-regulatory campaign, see the archive preserved by the Taiwan Association of Human Rights (2019): Taiwan Association of Human Rights (June 21, 2019). *Bushi xunxi fangzhi yezhe zilyu shijian zhunze [Code of conduct of self-regulation of misinformation prevention for practitioners]*. Taiwan Association of Human Rights. https://www.tahr.org.tw/sites/default/files/u87/190621_disinformation_code_of_practice_taiwan.pdf

14 For example, see the op-ed by the media scholar Yuan-Hui Hu (2019): Hu, Y. H. (2019, July 25). *Taiwan wangle pingtai zilyu da jiaxunxi, zhenneng fuhe shehuiqidai? [Can Taiwan's online platforms really meet society's expectations by regulating themselves to combat misinformation?]*. UDN Opinion. https://opinion.udn.com/opinion/story/12979/3948344

15 See the poll result conducted by the Taiwan FactCheck Center in 2022: Taiwan FactCheck Center. (June 19, 2022). *2022 jiaxunxi niandu dadiaocha: Taiwan shouci zhendui jiaxunxi xianxiang yu shishi chahe chengxiao da diaocha quanwen gongkai [2022 Annual Survey on Disinformation: Taiwan's first major survey on the disinformation phenomena and the effectiveness of fact-checking, full text published]*. Taiwan FactCheck Center. https://tfc-taiwan.org.tw/articles/7702

16 On the opaque content moderation practice, Legislator Michelle Lin criticized the moderation by Facebook during her interpellation at large, and required the Executive Branch to regulate such practice. The Premier at that time replied that "[the content moderation practice] shall not depart from and violate the national law … In this regard the NCC would study relevant laws so that to protect the rights of platform users." Liu, Y. C. (October 12, 2021). *Lianshu yanlun shencha bu touming, Su Zhenchang: Pingtai bude yuyue guonei fa [Facebook's content moderation is opaque. Su Tseng-chang: Platforms must not exceed domestic laws]*. Rti News. https://www.rti.org.tw/news/view/id/2113800 For the increasing presence of scams, see the news report on the investment scams disseminated on platforms: Chiu, C. F. (June 20, 2022) *Mao mingren touzi, you yuan fuweng meng… Jinnian qian 5 yue zha 11 yi, Lianshu, LINE zhan liucheng. [Impersonating celebrities for investment, luring individuals into dreams of becoming wealthy … Scamming 1.1 billion NTD in the first five months of this year, with occurrences on Facebook and LINE accounting for 60%]*. Liberty Times. https://news.ltn.com.tw/news/society/paper/1523964 For the questionable handling of user complaints, see the critique by Legislator Chia-yu Kao on the failure of handling user complaints by UberEats and Foodpanda: Hsu, S. Y. and Chiu, H. F. (May 19, 2020). *Candian chubao, kefu wuyong! Waisong pingtai zhengyi duo, lian Gao Jia-Yu dou shouhai. [Wrong orders and useless customer service! Food delivery platforms face multiple controversies, even Kao Chia-yu has been affected]*. FTV News Channel. https://www.ftvnews.com.tw/news/detail/2020519F02Q1

17 It is not entirely clear when the Government contemplated to integrate the disinformation-relevant procedure into the upcoming bill. For example, see Telecommunications Technology Center. (March 2021). *Yinying shuwei tongxun chuanbo fuwu fazhan zhi guiguan qushi yu fazhi gexin yanxi weituo yanjiu caigouan [Commissioned Research Project: A Study of Regulatory Trends and Legal Innovations in Response to the Development of Digital Communication Services]*. Telecommunications Technology Center. https://www.ncc.gov.tw/chinese/files/21042/5190_45998_210429_1.pdf. This report is the result of a comparative legal study project funded by the NCC. It shall be noted that this is the first government-related material that explored the possibility of inserting provisions that require platforms to display warning messages and that reiterated the need for court-issued access restriction of

harmful contents. However, a press release by the NCC about the upcoming draft of the bill (still using the name "Draft Bill of the Digital Communication Service Act") in December 2021 did not mention the above-mentioned procedures: NCC. (December 29, 2021). *NCC gongbu "shuwei tongxun chuanbo fuwu fa" cao'an jiagou, pan gongsi xieli gongtong jiangou anquan, kexinlai zhi wangle huanjing [NCC announced the draft framework of "Digital Communications Act," hoping that public and private efforts will jointly build a safe and reliable network environment]* [Press release]. https://www.ncc.gov.tw/chinese/news_detail.aspx?site_content_sn=8&cate=0&keyword=&is_history=0&pages=0&sn_f=46983

18 For example, see the interview to the chairman of TWIGF, Dr. Kuo-Wei Wu, and board member of the Judicial Reform Foundation Chun-Hung Lin: Huang, R. J. (July 6, 2022). *Shuwei zhongjie fuwufa zhuanjia you zhixing nandu gao [Experts worry about the difficulty of implementing the Digital Intermediary Service Act].* NOWnews. https://tw.news.yahoo.com/%E6%95%B8%E4%BD%8D%E4%B8%AD%E4%BB%8B%E6%9C%8D%E5%8B%99%E6%B3%95-%E5%B0%88%E5%AE%B6%E6%86%82%E5%9F%B7%E8%A1%8C%E9%9B%A3%E5%BA%A6%E9%AB%98-030002077.html Also see the report on the statement by the opposing Taiwan People's Party caucus: Pan, W. T. (July 22, 2022). *Pi "shuwei zhongjie fuwufa" lanquan ru "1450 taimianhua" minzhongdang yaoqiu zhongni cao'an [Taiwan People's Party criticized the "Digital Intermediary Services Act" for abusing power like "1450 coming to light" and demanded the NCC to redraft it*]. The Storm Media. https://www.storm.mg/article/4436745 and the op-ed of Legislator Kuei-Min Lee: Lee, K. M. (July 11, 2022). *Leekueimin xinsilu: dongzhujixian duanweilai - pingtai de yulun kongzhi [New thought of Kuei-Min Lee: Foreseeing and assessing the future - public opinion control of the platform].* Yam News. https://n.yam.com/Article/20220711272063

19 See the transcript (by civil society) of the public hearing aimed at platforms and trade associations, hosted on August 18, 2022: NCC. (August 18, 2022). *Pingtai fuwu yezhe yu gongxiehui [Platform Service Providers and Public Associations] [Public hearing transcript].* NCC. https://g0v.hackmd.io/@mrorz/ncc-disa/%2F%40mrorz%2FrJVVY2nC9

20 For one of the joint statements from three industry associations, see TiEA, DMA, & DEAT. (August 19, 2022). *"TiEA, DMA & DEAT sanda shuwei chanye xiehui lianhe shengming": shuwei zhongjie fuwufa yingyi shuwei fazhan wei mudi, jianqing zhanhuan lifa wu cangcu shanglu ["Joint Statement of three major digital industry associations - TiEA, DMA, and DEAT": The "Digital Intermediary Services Act" should be aimed at digital development, and it is suggested that the legislation should be paused and not rushed into operation.* https://drive.google.com/file/d/1jmTPZ7RfwJhG69KP8IGP7IaboPjHsrf6/view Also see the statement from the Asia Internet Coalition that was published after the halt of the rule-making process: Paine, J. (September 5, 2022). *Asia Internet Coalition (AIC) Comments and Recommendations on the Digital Intermediary Service Act (DISA),* Taiwan. Asia Internet Coalition. https://aicasia.org/download/229/

21 See the report of announcement: Lai, Y. C. (August 19, 2022). *Shuwei zhongjiefa zhanhuan gongtinghui, zhengyuan: rengxu duofang goutong, Su Tseng-chang chushou caishache [Public hearing on the suspension of the Digital Intermediary Service Act, Executive Yuan: communication is still needed, Su Tseng-chang stopped it].* CNA. https://www.cna.com.tw/news/aipl/202208190237.aspx. It is noted that such a decision was still seen as politically motivated by opposition parties and the general public, as the timing of the announcement was close to the upcoming local election, and the announcement was made by the premier not the NCC (supposedly politically independent). See the relevant news report: Tsau, Y. H. (August 24, 2022). *Zhengyuan dingtiao shuwei zhongjiefa ni hanka [The Executive Yuan has decided to pause the Digital Intermediary Service Act*]. Commercial Times. https://ctee.com.tw/news/policy/703444.html

22 Internal discussions with different governmental agencies and private exchanges with industry representatives. For example, one of the government agencies told the author that "it is imprudent to make DISA-like regulations just after its failure, so basically we ask platforms to voluntarily take down certain harmful contents based on their own terms of services." As far as the author is aware, similar requests were made by different governmental agencies on different harmful contents.

23 The coordinator in the Executive Branch on disinformation affairs, Minister without Portfolio Ping-Cheng Lo, explicitly elaborated in a forum hosted by Taiwan FactCheck Center on May 19, 2023, that the Government has knowingly adopted both approaches and has been waiting for the consensus required to propose another DISA-like framework legislation.

24 All these legislations were promulgated between February and June, 2023. It is also worth noting that the amendment to the Securities Trust and Investment Consulting Act, which tries to tackle investment fraud through misleading online advertisements, introduces joint liability for loss caused by illegal ads for platforms who failed to take down these ads within the time limit specified by the police. It is the first legislation other than copyright law that has similar designs in Taiwan.

25 While there were some lobbying activities, internet platforms and service providers were largely silent in public. Industry associations in some cases may issue statements, but their concerns never received public attention like those in DISA Draft controversy. The author was told, in a private exchange with platforms and service providers, that these legislations were narrowly focused to address social issues to which the public calls for solutions; thus, industry actors concluded that they would not receive much public support for open opposition to these legislations.

26 For example, Legislator Ching-Yi Lin has repeated this statement in various social media posts. Some of the aides of ruling party politicians, or online influencers who are ideologically aligned with the ruling party, also openly expressed similar views. The author also received the same feedback during private meetings with certain politicians.

27 For example, see the commentary by the TAHR: Chou, K. R. (August 29, 2022). *Shuwei zhongjie fuwufa cao'an zhuyao youquedian ni kandong le ma? [Do you understand the main advantages and disadvantages of the Digital Intermediary Services Act Draft?].* Taiwan Association for Human Rights. https://www.tahr.org.tw/news/3235 (which is the exception, there being very few public opinions from civil society sectors just after the onset of the controversies), and the scholar opinions in a DISA forum hosted by the local law media Plain Law Movement: Bai, T. Y. (January 17, 2023). *Yanlun ziyou yu wangle Anquan zai shuwei shidai zhong de liangnan – shuwei zhongjie fuwufa zhi xiankuang yu weilai (xia) [The Dilemma of Freedom of Speech and Internet Security in the Digital Era: The Present and Future of Digital Intermediary Services Law (Part 2)].* Plain Law Movement. https://plainlaw.me/posts/dilemma-of-disa2. For more negative views on the DISA Draft, see the argument by the Open Culture Foundation (OCF), another civil society group in Taiwan focusing on digital culture: Rock. (September 13, 2022). *Wanglu neirong shencha shei shuolesuan? [Who is in charge of internet content censorship?].* OCF. https://blog.ocf.tw/2022/09/blog-post.html; however, the OCF has also been critical about invasive personal data collection (for example, see its Ranking Digital Rights Project: https://ocf.tw/p/rdr/), though it did not actively express the relevant concerns in the controversies.

28 See another interview by Plain Law Movement: Bai, T. Y. (March 16, 2023). *Shuwei zhongjie fuwu tigongzhe de shidai zeren? Zhuanjia xuezhe guandian* [*The responsibility of the digital intermediary service providers? Viewpoints from Expert and Academic].* Plain Law Movement. https://plainlaw.me/posts/expert-opinion-for-disa; also see the other scholar opinions: Bai, T. Y. (January 17, 2023). *Yanlun ziyou yu wangle Anquan zai shuwei shidai zhong de liangnan – shuwei zhongjie fuwufa zhi xiankuang yu weilai (xia) [The Dilemma of Freedom of Speech and Internet Security in the Digital Era: The Present and Future of Digital Intermediary Services Law (Part 1)].* Plain Law Movement. https://plainlaw.me/posts/dilemma-of-disa1

29 There were a few panel and forum organizers who told the author that while they wish to organize discussion events on the DISA Draft, it had been very hard for them to invite panelists, as very few scholars wish to openly share their opinions. The author's own survey on relevant publicly available opinions also supports their observations.

30 See note 28, interview by Plain Law Movement.

31 For example, the transparency on the role of automatic decisions on content moderation and user complaint handling may give hints to abusers of inauthentic coordinated behaviors to finesse their tactics for information operations, which might conflict with the goal of tackling online disinformation.

32 For the rough definition of integrity workers, the author find the explanation from the Integrity Institute quite useful: https://integrityinstitute.org/what-is-an-integrity-worker.

33 The author, however, does not endorse complete and direct legal transplantation of legislations from leaders of regulations, as contexts and policy goal priorities differ.

> The failed rule-making process of the DISA Draft did initiate the first formal discussion in Taiwanese society regarding the need to hold platforms accountable and has become a (potentially negative) reference for future policy development. The controversy clearly demonstrates the limitations of a traditional top-down approach to rule-making, where the government determines the agenda without sufficient consensus from various stakeholders and the general public, thus eroding the mutual trust.
>
> **Anonymous**

STRIKING A BALANCE: CONTENT MODERATION AND FREEDOM OF EXPRESSION IN LATIN AMERICA

Priscilla Ruiz Guillén

The controversial issue of whether to regulate digital platforms is an international-level debate due to the so-called Big Tech gatekeepers' position regarding content.[1] In terms of content moderation, these digital platforms' economic and discretionary power is often the spotlight of public debates. This situation has raised various discussions in Latin America, especially in countries such as Mexico, Colombia, Argentina, and Brazil.

Over the past five years, there has been a constant call to regulate social media platforms in Latin America, mainly due to the incongruence of how these platforms have moderated the content not only of individuals but also of governmental officials who applied these platforms to disseminate their opinions, programs, discussions, or criticisms. On the other hand, civil society, such as activists and human rights defenders, also uses the platform to raise awareness about issues that are of interest to their audience,[2] including LGBTQ+, feminists, and migrants, but calls for regulations often aim to stifle the voices of these groups. For example, feminist collectives who have been using social media platforms to draw attention to the issue of violence against women, girls, and adolescents are often silenced by community standards and norms when they share information about sexual and reproductive rights, including abortion. Another example are the journalists and human rights defenders who use these platforms to shed light on severe human rights violations within their countries as well as issues of corruption and money laundering within government institutions.[3]

The European Union (EU) stands out as one of the leaders in promoting legislation regulating content moderation and protecting personal data stored by large digital platforms. These regulations have prompted discussions to promote transparency in how algorithms work regarding content moderation. In Latin America the approach towards law enforcement for content moderation on social media platforms has been different. Adopting the EU initiatives in Latin America could result in unintended consequences that would undermine human rights in the digital space, as has been documented.

The analysis of Latin America requires a deep understanding of its diverse socioeconomic and political landscape. Several countries are considered consolidated democracies, while others experience political regimes that reluctantly suppress protection and respect for human rights and democracy. Over the past five years, a wide range of legislative proposals aimed at regulating digital platforms have emerged, with the following noteworthy trends:

1. Some initiatives seek to undermine the voices of dissidents and marginalized groups.
2. Various regulations introduced generic, ambiguous, unclear, and subjective definitions of hate speech, sexism, and misogyny. These regulations also increased state surveillance of content published or shared on social media.
3. Several regulatory proposals focused on imposing excessive administrative sanctions solely on major digital

platforms like Big Techs, leaving out small platforms without economic dominance.

4. A lack of consensus can be seen regarding the understanding and application of content moderation, since major platforms like Google, Twitter, Meta, and TikTok each adopt a different self-regulation model.

In the following sections, this article will introduce notable case studies from Latin America, shedding light on the various regulatory challenges and their implications.

Brazil and the "Fake News Bill" 2630/2020

In 2014 Brazilian congress adopted a law known as the *Marco Civil do Internet*,[4] which stated that companies are not held liable for content published by third parties unless they fail to comply with a court order to remove it or when the content contains nonconsensual nudity. The *Marco Civil do Internet* was a global benchmark and was recognized by the Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights as a "milestone in the advancement of the regulatory framework in relation to the issue of freedom of expression and privacy in the digital space"[5]. Then, Provisional Measure No. 1068[6] was approved, issued on September 6, 2021 by the Executive Power of Brazil, which modifies the Internet Civil Framework and the Copyright Law to regulate the use of social networks in Brazil. This Provisional Measure No. 1068 was criticized by several civil society organizations, since it introduced a series of provisions that allowed companies to remove content or suspend accounts. These actions could be carried out without a court order only for "just cause." Later in 2021, under the administration of Jair Bolsonaro, the Fake News Act, known as Act 2630/2020, was enacted with immediate effect. The main goal was to combat the arbitrary deletion of accounts, profiles, and online content. The intention of Act 2630/2020 was to clarify policies of social media platforms and offer users the opportunity to republish content that had been removed based on alleged violations of platform policies or community standards. This initiative came after Bolsonaro himself had content removed for spreading false information about COVID-19, which he considered censorship.[7] It is noteworthy that discussions at that time around the so-called fake news excluded civil society organizations and other civic actors who wanted to participate in the process and issue recommendations to make sure this bill could meet the standards of human rights and freedom of expression.

The debate on content moderation of the Act 2630/2020[8] was reignited after supporters of the former Brazilian president Jair Bolsonaro stormed the headquarters of the three branches of government in Brasilia on January 8, 2023. The act was seemingly incited by disinformation on social media claiming that the actual president, Lula da Silva, had fraudulently lost the reelection, and thus Bolsonaro should be the real winner of the election.

Considering these events, the possibility of re-discussing Act 2630/2020 has emerged with support from the new government administration, the judiciary, and a few civil society groups[9]. Act 2630/2020, described above, has been controversial since its inception and remains a topic of contention. It not only challenges the power of Big Tech but also raises concerns about its broad, inaccurate, and vague definition of law concepts and implementations that could restrict online freedom of expression in Brazil. The approval of this law has drawn a fine line between different civil society groups. On the one hand, there are those who support the law, advocating for transparency, accountability, and a clear appeal process to address arbitrary content moderation by social media platforms. On the other hand, some argue that placing the responsibility on intermediaries to remove, block, filter, and monitor "unacceptable" content according to the law's mandate should not be allowed, as the subjective interpretation by the authority could lead to censorship of legitimate expression.

Mexico: Attempts to Censor the Digital Sphere

In 2021, Senator Ricardo Monreal, a member of the majority party MORENA, introduced a legal proposal to regulate social media platforms and established conditions for state intervention. The proposal sought to modify the Federal Law of Telecommunications and Broadcasting and provide powers to the Federal Institute of Telecommunications to oversee and dictate the actions of social media platforms.

The discussion oversees the necessity of limiting speech in the digital realm, particularly its compliance with international human rights standards related to freedom of expression (FoE). A legal proposal published by the Article 19 Regional Office for Mexico and Central America flagged a series of attempts against FoE which include:[10] (i) previous censorship by any governmental authority that under their belief and interpretation found that any content published, distributed, or commented should be removed immediately without considering the due process of law principles; (ii) leaving an open door to surveillance by government authorities of any content considered harmful or that jeopardizes national security issues or any other activities that were not considered appropriate to society; and finally (iii) authorities could pressure companies to censor content or delete the accounts of political opponents, activists, journalists, and human rights defenders who express views that are not aligned with the current official narrative.

Since 2021, civil society organizations in Mexico have been advocating for an open and informed discussion on internet regulations to ensure that possible solutions would safeguard the exercise of freedom of expression and human rights in the digital sphere rather than restricting them. The proposal to modify the Federal Law of Telecommunications and Broadcasting and provide powers to the Federal Institute of Telecommunications was withdrawn, and the debate regarding a new one has not been raised. Nevertheless, Mexico has seen other content moderation initiatives involving digital platforms, such as copyright issues outlined in the Federal Copyright Act, tackling cases related to content moderation and removal of nonconsensual images, which are sanctioned under various penal codes at both the state and federal level. On the other hand, there are efforts to formulate laws on digital market and audiovisual services; regulations specifically tailored to new digital platforms within the telecommunications act have also been raised.

One of the main challenges in Mexico lies in the persistence of government positions that aim to regulate the digital sphere without understanding its complexities and relevant civil rights. Furthermore, there is a constant lack of transparency and limited participation in the policy-making process, excluding civil society from this process. Finally, the authorities in Mexico tend to promote a "techno solutionism" approach, which seeks to simplify solutions to complex problems within the country.

Argentina and Agreements on the Rights of Platform Users

The focus of digital platform regulations in Argentina shifts toward fiscal measures targeting service platforms under two premises: on their industry profile and on their impact on cultural consumption. In addition, one of the proposed attempts to regulate content moderation is based on copyright cases.

However, between 2021 and 2022, a series of meetings took place in Argentina to tackle the content moderation issue carried out by digital platforms and, above all, its impact on users' human rights. These dialogues were mainly led by civil society organizations and resulted in the "Agreements on the rights of users of platforms in Argentina before the moderation of content,"[11] released in November 2022.

These agreements contribute to the debates on the rights of platform users and provide suggestions for legal proposals. Civil society organizations emphasize the compatibility with freedom of expression and the guiding principles of business and human rights, highlighting that first, content moderation should be compatible with the framework of international human rights law and Inter-American standards of freedom of expression. Second, regulatory frameworks need to comply with strict standards of necessity, proportionality, and legality. Third, potential regulations must not be in conflict with international human rights law. Fourth, protection of freedom of expression and the rights of users of platforms that are affected by content moderation should be the priority.

Moreover, the agreements outline actions to implement the proposals and recommend the establishment of new or improved mechanisms for social media users. While there are currently no bill proposals aimed at regulating big tech companies for content moderation, it is important to note that the agreements prioritize principles such as due process and transparency in the criteria used for moderation, particularly in relation to the use of algorithms. Additionally, the agreements seek to foster a multi-stakeholder dialogue to urge policymakers to adopt the international human right standards in future attempts to regulate the digital realm.

Colombia: Regulations to Protect Children's Rights

The Colombian government has focused more on regulating content for the protection of children's rights by pressuring media outlets and internet service providers (ISPs) regarding the dissemination of potentially harmful content. This initiative, known as Bill 600 (PL 600), is in the First Commission of the Colombian House of Representatives, where it passed the first stage of approval. Meanwhile, there are concerns regarding its potential constraints on freedom of expression and the possibility of activating mechanisms of prior censorship.

The PL 600 bill intends to safeguard children from media responsible for the content they disseminate, especially within social media platforms, and holds ISPs responsible for any content that violates the guidelines on violent or sexual content, as well as the categorization of content in broadcasting programs. In addition, the PL 600 proposes that ISPs implement internal security systems to avoid unauthorized access to their networks and counter acts that violate children's rights. For example, the bill includes a series of prohibitions ranging from hosting images, texts, documents, or audiovisual files that harm the psychological well-being and integrity of children. Moreover, the bill stipulates that internet providers must report any criminal acts affecting children by disseminating harmful or inappropriate content.

Civil society organizations such as the Foundation for the Freedom of the Press, worried that the bill promotes censorship and restricts media freedom to disseminate their own editorial line. They argued that this bill would grant the

authority power to limit content based on subjective interpretation of what is defined as violent or harmful.

Central America: Honduras and El Salvador Bills Increasing Penalties

Recent legal attempts in El Salvador and Honduras have raised concerns, as they appear to undermine the voices of dissidents and vulnerable groups that rely on social media platforms for legally protected expressions.[12] In both cases, there are regulations proposed to increase penalties for defamation, slander, and libel: The latter two specifically target digital media outlets.

In 2019 in El Salvador, there was an effort to reform the Penal Code that attempted to criminalize defamation, slander, libel, and the dissemination of false information using false profiles.[13] This legal attempt seeks to punish acts that harm an individual's honor, personal privacy, and self-image. The punishment for these offenses includes imprisonment for 4 to 8 years.

Similarly, in Honduras, amendments were made to different legal frameworks, including the penal code of the country, which increased the crime of slander and libel when committed through collective disclosure websites or social media.[14] The proposal to amend the Penal Code was approved and made official on November 10, 2019. It also introduces the concept of "direct defamation." According to the regulation, publication, reproduction, repetition of slander or defamation imputed by another, and making accusations in an impersonal way or using similar terms are by definition "direct defamation."[15]

These regulatory efforts, aimed at addressing online behavior, portray the diminishment of democracy and human rights in two countries of Central America, where civil society organizations have less support than in other countries of the region. Initiatives in these countries have raised concerns about the concentration of power in the case of El Salvador. Many local civil society organizations worry that the use of this legal framework will be a perfect excuse to criminalize dissident voices that rely on digital platforms to express themselves and disseminate information. Critics argue that these laws may be used as a tool to inhibit journalistic practice in corruption investigations among governmental officials, members of the security forces, or private actors

Transparency and Due Process in Content Moderation at the Hands of Civil Society in Latin America

Transparency and due process in content moderation by digital platforms have been major concerns globally highlighted by civil society organizations. In 2021, the Inter-American Commission on Human Rights (IACHR), through the Special Rapporteur on Human Rights (RELE), fostered the Dialogue of the Americas, aiming to address the challenges in the Latin America region. As a result of several consultations, the RELE summarized three main topics that need to be addressed while considering the coexistence of democratic values and human rights in digital spaces:

1. Deterioration of public debate: Since public debate in digital spaces often lacks institutional or social oversight, online conversations often become uninformed and disrespectful, offering greater spotlights to voices that promote racism, discrimination, and disinformation. This poses a threat to democratic values.
2. Digital literacy deficit: Countering new challenges, such as disinformation or anti-democratic speech, requires an empowered society that needs the tools to access information amid authoritarian governments. There is a need to invest in promoting digital literacy.
3. Content moderation: The region faces challenges resulting from content moderation and should be protected by the Inter-American human rights framework.

The level of multi-stakeholders' participation plays a crucial role in promoting an inter-American agenda that tackles these challenges, as inclusive participation can foster dialogue and present solutions that are viable, feasible, and contextually relevant to each Latin American country. However, it has been observed that governments within the region have limited to no participation in these dialogues. Of the 30 members of the Organization of American States, only 11 have participated and 6 have provided input. Moreover, only 3 countries have the sustained capacity or resources to continue with the inter-American strategy proposed by the IACHR, especially in governance forums relating to the internet.

Conclusion

In Latin America, as in other regions of the world, not all platforms operate under the same content moderation framework as the dominant Big Tech. For instance, the Latin America Internet Association, an association representing the Big Tech agenda, explains that many of the regulations they supported are related to personal data protection, economic competition, consumer protection, intellectual property protection, and the streaming of sporting events, among others.

Meanwhile, platforms such as Wikimedia, a nonprofit organization that hosts Wikipedia, adopt a horizontal model where users are actively involved in decision-making pro-

cesses regarding content and behavioral policies. This bottom-up approach aims to build trust and foster community participation.

Understanding the business models and nature of platforms is fundamental to grasping their content moderation practices. Profit-seeking companies often have a top-down decision-making model, whereas nonprofit companies prefer a bottom-up decision-making process on content moderation, supporting the internet conception as free in its interoperability to sustain the architecture of the internet.

Current regulations, for example the "Safeguarding freedom of expression and access to information: guidelines for a multi-stakeholder approach in the context of regulating digital platforms"[16], discussed in the United Nations Educational, Scientific and Cultural Organization (UNESCO), often overlook platforms like Wikimedia who prioritize alternative content moderation models, and instead focus solely on regulating Big Tech and its market dominance.

In order not to fall unduly in the existence of these models that have a great social impact, authorities should focus on regulating digital platforms under a market dominance and competition law to disempowerment of the communities that decide on the moderation and curation of their content to make it a reliable space.

When considering attempts to regulate digital platforms, it is important to define the precise scope and nature of the content that needs to be regulated while also ensuring that human rights in the digital space are not undermined. Authorities should uphold international standards regarding human rights when addressing issues such as disinformation, hate speech, sexist speech, or any other form of expression. In addition, policymakers should take into account the cultural and sociopolitical context of individual countries to avoid the undue restriction of legitimate content. Therefore, it is essential to maintain an ongoing dialogue and incorporate multi-stakeholder participation at all levels of discussion, thereby preventing the deterioration of public debate. Finally, it is important to recognize that relying solely on "techno-solutionism" and regulations may not always provide solutions to complex issues. Instead, a shared responsibility of both society and the government needs to be explored and undertaken as part of a broader regional agenda.

References

1 Catalyst For Collaboration. Content Moderation. June 2020. https://catalystsforcollaboration.org/blog/

2 Article 19 Mexico and Central America Regional Office. January–March 2023. *Barometro de la Libertad de expresiñon en Centroamerica y Cuba*. https://articulo19.org/barometro-de-la-libertad-de-expresion-en-centroamerica-y-cuba-enero-marzo-2023/

3 Article 19 Mexico and Central America Regional Office. January 15, 2022. *Protesta digital: Una transformacion histórica de la libertad de expresión en el mundo*. https://articulo19.org/protesta-digital-una-transformacion-historica-de-la-libertad-de-expresion-en-el-mundo/

4 Coalizao Direitos Na Rede. Por una regulación participativa y responsable. April 14, 2021. https://direitosnarede.org.br/2023/04/14/firma-la-peticion-por-una-regulacion-de-ia-participativa-y-responsable/

5 Relatoria Especial para la Libertad de Expresión. Comunicado de Prensa R237/21. La Relatoría advierte sobre los riesgos para el derecho a la libertad de expresión en internet en Brasil frente a la reforma del Marco Civil de Internet. September 9, 2021. https://www.oas.org/es/cidh/expresion/showarticle.asp?IID=2&artID=1210#:~:text=El%20Marco%20Civil%20de%20Internet%20de%20Brasil%2C%20adoptado%20en%202014,Am%C3%A9ricas%20como%20en%20el%20mundo

6 Diario Oficial da Uniao. Medida Provisória N. 1.068. September 6, 2021. https://www.in.gov.br/en/web/dou/-/medida-provisoria-n-1.068-de-6-de-setembro-de-2021-343277275

7 RTVE spanish. *Bolsonaro cambia regulaciones a regulación de las redes sociales para que no puedan eliminar cuentas arbitrariamente*. September 7, 2021. https://www.rtve.es/noticias/20210907/bolsonaro-cambia-regulacion-redes-sociales/2169055.shtml

8 France 24. *El proyecto para regular redes sociales que causas polémica en Brasil*. May 12, 2023. https://www.france24.com/es/minuto-a-minuto/20230512-el-proyecto-para-regular-redes-sociales-que-causa-pol%C3%A9mica-en-brasil

9 Observacom. *Cuáles son las propuestas del gobierno de brasil para regular las plataformas de contenido*. April 6, 2023. https://www.observacom.org/cuales-son-las-propuestas-del-gobierno-de-brasil-para-regular-las-plataformas-digitales-de-contenido/

10 Cortés Roshdestvensky, Vladimir, y Martha A. Tudón M., "*Coordenadas para el análisis: Trump y las plataformas digitales", Animal Político*, January 18, 2021. https://www.animalpolitico.com/altoparlante/coordenadas-para-el-analisis-trump-y-las-plataformas-digitales/

11 Access Now. *Argentina: Firma acuerdos sobre derechos ante moderación de contenidos*. November 14, 2022. https://www.accessnow.org/press-release/argentina-acuerdos-derechos-en-plataformas-ante-moderacion-contenidos/

12 Alianza Regional. *Internet, Libertad de Expresión y Espacio Cívico en América Latina*. 2021. https://www.alianzaregional.net/wp-content/uploads/2021/05/Articulo-XIII-1.pdf

13 El Mundo: Diario Libre y Objetivo. *Diputados revivirán propuesta para castigar difamaciones e injurias en redes sociales*. July 21, 2020. https://diario.elmundo.sv/Pol%C3%ADtica/diputados-revivirian-propuesta-para-castigar-difamaciones-e-injurias-en-redes-sociales

14 ARTICLE 19 Mexico and Central America. *Honduras: Nuevo Código Penal exhibe a un Estado que criminaliza la libertad de expresión y el acceso a la información*. May 2, 2020. https://articulo19.org/honduras-nuevo-codigo-penal-exhibe-a-un-estado-que-criminaliza-la-libertad-de-expresion-y-el-acceso-a-la-informacion/

15 Access Now. *Derecho al honor vs. derecho a la Libertad de expresión: regulación de contenidos en Perú, El Salvador y Honduras*. June 28, 2019. https://www.accessnow.org/derecho-al-honor-vs-derecho-a-la-libertad-de-expresion-regulacion-de-contenidos-en-peru-el-salvador-y-honduras/

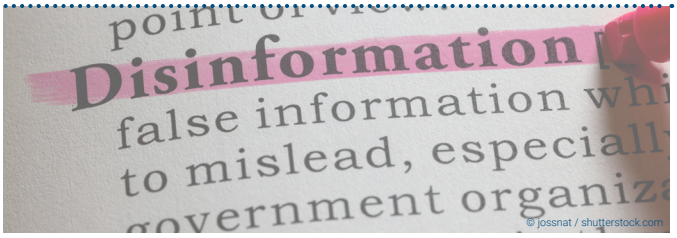
16 UNESCO, CI-FEJ/FOEO/3 Rev., Safeguarding freedom of expression and access to information: guidelines for a multistakeholder approach in the context of regulating digital platforms. Conference: Internet for Trust - Towards Guidelines for Regulating Digital Platforms for Information as a Public Good, Paris, 2023.

The level of multi-stakeholders' participation plays a crucial role in promoting an inter-American agenda that tackles these challenges, as inclusive participation can foster dialogue and present solutions that are viable, feasible, and contextually relevant to each Latin American country. However, it has been observed that governments within the region have limited to no participation in these dialogues.

Priscilla Ruiz Guillén

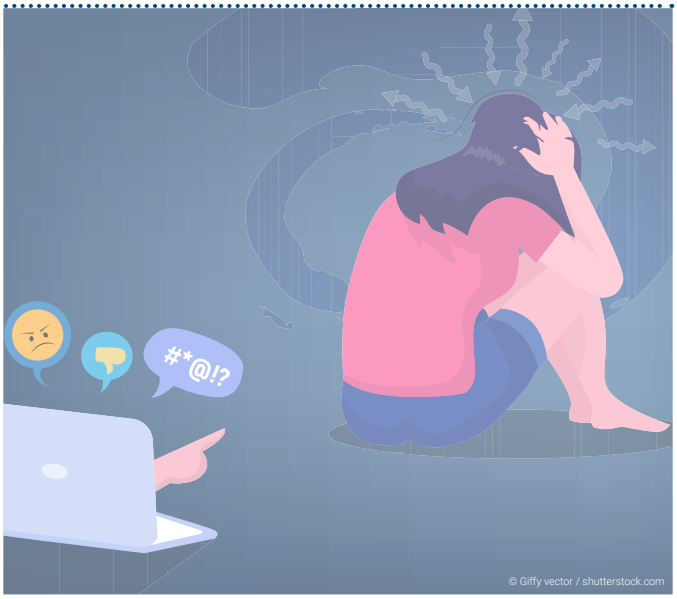
ABOUT THE AUTHORS

Alexander Hohlfeld is a digital policy expert who analyzes the role of digital platforms in social processes as well as in regulatory and educational approaches. He currently specifically focuses on the analysis of approaches to evaluate recommender systems of internet intermediaries.



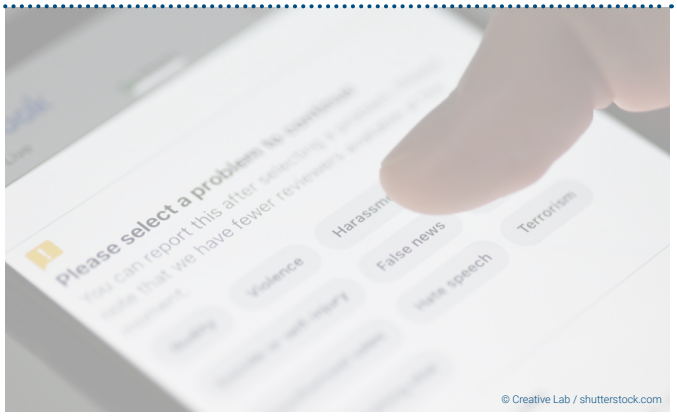
Alphonce Shiundu is an award-winning writer and journalist and the pioneer country editor of Africa's leading fact-checking organization, Africa Check. He previously reported on public policy, legislation, and politics for Kenya's mainstream media, specifically The Nation Media Group and The Standard Group Plc. Alphonce's work at Africa Check has garnered international recognition, with reports from prominent media outlets such as the BBC, CNN, Aljazeera, the Financial Times, and the New York Times. He has also conducted dozens of workshops on evidence-based decision-making, countering disinformation, access to information, media literacy, and fact-checking for government officials, researchers, journalists, students, and civil society members in Africa and Europe.

Ann Cathrin Riedel is the managing director of NExT e.V. and chairwoman of LOAD e.V. - Association for Liberal Internet Policy. Previously, she was responsible for international digital policy at the Friedrich Naumann Foundation for Freedom. Ann Cathrin holds advisory roles in various governmental agencies, including the Digital Advisory Board for the Federal Ministry for Digitalization and Transport and the Digital Council of the State of Saxony-Anhalt. She was also an expert on digital policy issues in the German Bundestag and Berlin House of Representatives. In 2020, Capital magazine awarded Ann Cathrin as one of the "Top 40 under 40" in the Science and Society category. Through her publications, speeches, and the award-winning newsletter "Ann Cathrin's Digital Digest," she explores topics such as ethics, civil rights, freedom of expression, communication in the digital space, and digital sovereignty.



Dr. Gehan Gunatilleke is a lawyer specializing in constitutional law and international human rights law. He is a founding partner at LexAG, a law firm specializing in Sri Lankan civil and public law, and a junior research fellow at Pembroke College, University of Oxford.

Chung Ching Kwong is a political and digital rights activist from Hong Kong. Previously, she was the spokesperson for Keyboard Frontline, an organization dedicated to monitoring censorship and digital rights, and a columnist at Apple Daily. Chung Ching first got involved in activism in 2012 due to the Copyright Amendment Bill in Hong Kong. She is currently pursuing her PhD in law at the University of Hamburg, focusing on data protection laws. She is currently the Hong Kong Campaigns Coordinator at the Inter-Parliamentary Alliance on China (IPAC) and a columnist at Welt am Sonntag in Germany.



Mu-Huan Wang is a senior research fellow at the Telecom Technology Center (TTC), where he serves as a digital media policy expert, leading programs and research efforts to support national communications and internet regulation. With over 12 years of experience in Taiwanese think tanks, he has provided legal support across multiple government ministries. Before joining the TTC, he was a manager at the Science & Technology Law Institute of the Institute for Information Industry, formulating the Digital Convergence Development Program for the Executive Yuan. Currently pursuing a PhD in law at National Chengchi University, Mr. Wang holds an M.I. in Social Informatics from Yuan Ze University and an LLB from National Taipei University.

Priscilla Ruiz Guillén is a human rights lawyer and legal coordinator for the Digital Rights Program at Article 19 Office for Mexico and Central America. With expertise in strategic litigation, she has worked as a consultant with organizations such as Amnesty International and the Inter-American Commission on Human Rights, advocating for freedom of expression and digital rights. Priscilla has contributed to reports and hearings at the United Nations and other international organizations, focusing on online freedom of expression, artificial intelligence, content moderation, and online protests in Mexico.