White Paper on the Framework for an Online Harms Protection Bill in Nigeria

THE SECRETARIAT OF THE ONLINE HARMS PROTECTION BILL PROJECT

Advocacy for Policy and Innovation (API) in Partnership with the
National Information Technology Development Agency (NITDA)

DECEMBER 2024

Table of Contents

Executive Summary

Every online action connects individuals globally, fostering opportunities but also exposing vulnerabilities. Social networks, educational platforms, community fora, and online games have become the new public squares where ideas and knowledge are exchanged every minute at breathtaking speeds. These connections are achieved online to power new services while adding value and efficiency to systems and processes.

Yet, an emerging challenge of various types of harm exists within the cause-and-effect of this bustling digital commons and marketplace. These online harms threaten the emerging online architecture and adversely affect all facets of human life.

Online harm has stretched its reach into every corner of the globe, and Nigeria's lively digital community is also in the grip of these threats. Cyberbullying, the rising wave of hate speech and extremism, and the proliferation of misinformation and disinformation colloquially referred to as "fake news" demonstrate how complicated our online world has become.

A fundamental duty of a state is the preservation of the rights of its citizens, including digital rights and the protection of these citizens from all categories of harmful incidents. Similarly, businesses have a duty and responsibility to protect rights, too, and states have a duty to ensure policies enable businesses to protect rights. Addressing the challenge of online harms has led to a steady issuance of new laws and rules in Nigeria, such as the the Internet Code of Practice by the Nigerian Communications Commission 2017, Nigeria Broadcasting Code 2020, and the National Information Technology Development Agency (NITDA) issued Code of Practice for Interactive Computer Service Platforms/ Internet Intermediaries, 2022 . Legislative and regulatory proposals such as the *Digital Rights and Freedom Bill 2019,* the *"Social Media Bill,"* and the *NBC Amendment Bill 2023* aim to shield citizens from these online dangers. These efforts are intended to create a regime for intermediary liability and a framework for digital content moderation (CM) in Nigeria, emphasising responsibility for internet platforms and intermediaries operating in the country. These regulatory efforts have encountered criticism mostly around perceived inadequacies, the constriction of the civic space, and constraints on human rights. This is even as these online threats keep evolving and the need to balance human rights and civic protections under legal frameworks becomes more apparent.

"Online harm has stretched its reach into every corner of the globe, and Nigeria's lively digital community is also in the grip of these threats.…..This underscores the urgency of addressing online harms in Nigeria"

Global experience indicates that CM-centred approaches, whether human-based or augmented by machines, have inherent limitations that have necessitated several shifts in strategy. Nigeria must, therefore, consider its landscape and realities to craft new, fair and effective rules.

This white paper proposes a shift from the current CM-centred approach, leaning on a patchwork of laws and rules in Nigeria that essentially grants the platforms responsibility for monitoring content, towards a coherent, coordinated framework that guarantees citizens' rights while shielding society from the harms of the internet.

CM possesses its merits, but its increasingly exclusive prioritisation as government intervention's purpose and end state must correctly capture the digital space's dynamism and the evolving complexities of harmful online content. Regulations will constantly engage in a losing game of catchup with information technology-driven innovation and, ultimately, with the abuses of these innovations generally and online harm in particular. This is because human-based systems, susceptible to bias, and machine-augmented processes have struggled to achieve nuanced contextual understandings of data. Also, the subjective nature of what may constitute harmful content and the legal and social differences within societies that share similar platforms create a considerable challenge in determining this type of content. Therefore, the future lies in adopting a system for a mutual understanding of the landscape of online harms, establishing a "duty-of-care" proposition, and adopting a stakeholder-led approach.

These will not replace CM practice but enhance it to improve the protection of rights and encourage proactive action to prevent the abuse of information technology. At this proposed model's core lies a co-regulatory approach that includes civil society participation, rules-obligating platforms, and transparency mechanisms for citizen involvement.

This paradigm shift emphasises improving transparency in intermediary liability processes and establishing clear compliance measures. It proposes a strategic evolution to a *"duty-of-care"* ethos, stakeholder partnership, and coordination that fortifies societal and national defences while qualitatively elevating the present approach from content moderation simpliciter to online harms protection. This paper proposes a digital landscape where safety and rights coexist under a draft Online Harms Protection Bill (OHP Bill) to be submitted for legislative scrutiny. Under the proposed regime, it is intended that transparency will be the beacon guiding a national strategy to regulate third-party content in Nigeria. Once enacted, the framework will apply to all online platforms in Nigeria.

In light of these propositions, this white paper:

- identifies key gaps in Nigeria's digital governance frameworks and proposes a multi-stakeholder approach to tackling online harms

- presents a comprehensive national framework that outlines specific responsibilities for public and online platforms, including establishing transparent procedures for addressing harmful content and imposing penalties for non-compliance.

- calls for instituting a coordination and research centre that will work with stakeholders to coordinate the implementation of the national framework, curate Nigeria's journey in protecting and enabling online content, and promote digital rights in Nigeria.

List of Abbreviations

CM	Content Moderation
NBC	National Broadcasting Commission
NCC	Nigerian Communications Commission
NHRC	National Human Rights Commission
NITDA	National Information Technology Development Agency
UGC	User Generated Content
OCSEA	Online Child Sexual Exploitation and Abuse
OHP	Online Harm Protection
ONSA	Office of the National Security Adviser
OSP	Online Service Providers
CSEAM	Circulation of Child Sexual Exploitation and Abuse Materials

Acknowledgements

The development of this white paper on Nigeria's Online Harms Protection (OHP) framework was made possible through the collaborative efforts of a diverse group of individuals and organisations who brought their expertise, insights, and dedication to this critical project. We extend our heartfelt gratitude to all who contributed to creating a framework to enhance digital safety and foster a more inclusive and secure online environment in Nigeria. We particularly acknowledge the partnership and effort of Mr Kashifu Inuwa, the Director General of the National Information Technology Development Agency (NITDA), for his unwavering support and commitment to ensuring the conversation on online harms in Nigeria is stakeholder-led.

We acknowledge and thank the following groups for their invaluable contributions:

1. Institutional Partners
Through their membership in the steering team, the National Information Technology Development Agency (NITDA) and the National Human Rights Council (NHRC) provided critical institutional guidance. Their expertise and insight were instrumental in ensuring the framework's feasibility and technical robustness.

2. Secretariat
The secretariat for developing this white paper was responsible for the heavy lifting of the research, editing, and articulating the issues covered. The white paper's pages reflect the team's expertise, hard work, and research effort. We acknowledge the willingness to share resources and the sacrifice to write, read, and rewrite the paper painstakingly. We recognise the efforts of Tech Hive Advisory and Tech Policy Advisory as organisational members of the secretariat who dedicated resources and time to support the effort.

3. Steering Team
The steering team's expertise, oversight, counsel, and strategic input were crucial throughout the project. Their leadership and support enabled a cohesive, wellstructured approach to the development process, ensuring that the framework reflected diverse perspectives and aligned with national and international standards.

4. Stakeholders and Community Representatives
This project benefited from the active participation of a wide array of stakeholders, including civil society organisations, representatives from the private sector, government agencies, digital rights advocates, academic institutions, and members of the public. We especially thank the members of vulnerable communities who shared their experiences and insights, allowing us to tailor the framework to the needs of those most affected by online harms.

5. Partner Organisations
Special thanks to our partner organisations who supported this project by facilitating workshops, conducting research, and providing resources and guidance. Your commitment to digital rights and safety in Nigeria has been invaluable in shaping this framework. Specifically, we thank Paradigm Initiative for providing focused platforms to socialise this effort and provide opportunity for valuable feedback and input.

6. Advisors and External Reviewers
Our gratitude goes out to the advisors and reviewers who offered their time and expertise to provide feedback and recommendations on the draft versions of this framework. Your insights ensured the final document was well-informed, balanced, and actionable. We specifically thank Mr Richard Ali for graciously editing this paper.

We have reserved space here to name each member of the Steering Team and Secretariat whose contributions were essential to this project's success.
List of Stakeholders and Contributors:

Steering Team

S/N	Names	Organisation
1	Kasim Sodangi	Advocacy for Policy and Innovation (API)
2	Victoria Manya	Advocacy for Policy and Innovation (API)
3	Dr Aminu Lawal	National Information Technology Development Agency (NITDA)
4	Abdullahi Aliyu	National Information Technology Development Agency (NITDA)
5	Bashira Hassan	National Information Technology Development Agency (NITDA)
6	Gbenga Sesan	Paradigm Initiative
7	Ridwan Oloyede	Tech Hive Advisory
8	Moses Faya	Tech Policy Advisory
9	Rosemary Ajayi	Digital Africa Research Lab
10	Catharine Angai	PhD Researcher, University of Sussex
11	Dr Tomiwa Ilori	University of Pretoria
12	Adegoke Adeboye	Paradigm Initiative
13	Saadatu Falii Hamu	Hamu Legal
14	Rotimi Ogunyemi	BOC Legal
15	Dr Adekemi Omotubora	University of Lagos
16	Chioma Agwuegbo	Tech Her
17	Dawn Dimowo	Policy Expert
18	Tope Ogundipe	Tech Societal
19	Chinenye Uwanaka	The Firma Advisory
20	Judith Manjel Jabbe	National Human Rights Commission (NHRC)
21	Adamu Halilu	National Human Rights Commission (NHRC)

Secretariat Members

S/N	Names	Organisation
1	Kasim Sodangi	Advocacy for Policy and Innovation (API)
2	Victoria Manya	Advocacy for Policy and Innovation (API)
3	Abra Dangnan	Advocacy for Policy and Innovation (API)
4	Veronica Pana	Advocacy for Policy and Innovation (API)
5	Donald Sikpi	Advocacy for Policy and Innovation (API)
6	Esther Makaino	Advocacy for Policy and Innovation (API)
7	Stella Yakubu	Advocacy for Policy and Innovation (API)
8	Abdullahi Aliyu	National Information Technology Development Agency (NITDA)
9	Moses Faya	Tech Policy Advisory
10	Ridwan Oloyede	Tech Hive Advisory
11	Dorcas Tsebee	Tech Hive Advisory
12	Diana Uzor	Tech Hive Advisory

Caveat

This white paper is intended to provide insights, research findings, and recommendations for developing a framework for Nigeria's Online Harms Protection (OHP) Bill. The document reflects the authors' and contributors' collective knowledge, expertise, and opinions at publication. It also represents the research and experience of authors while engaging various stakeholders in Nigeria over the past two years. However, it does not constitute legal advice or replace formal guidance from regulatory or legislative bodies.

The content herein is based on the information available, analysis conducted, and stakeholder feedback obtained during the white paper's preparation. As the digital landscape and regulatory environment continue to evolve, the recommendations and conclusions in this document may be subject to change. Stakeholders, policymakers, and readers are encouraged to verify any information independently and to consider local legal and regulatory requirements before acting on the recommendations provided.

While every effort has been made to ensure the accuracy of the information presented, the authors and contributors accept no responsibility or liability for any actions taken based on this white paper. Additionally, references to specific frameworks, organisations, or legal instruments are provided for informational purposes only and do not imply endorsement or affiliation.

This white paper is intended as a foundational document to foster discussion, collaboration, and further research on online harm protection in Nigeria. It is not intended to serve as a final or legally binding document.

Methodology for Developing the Framework for Nigeria's Online Harms Protection Bill

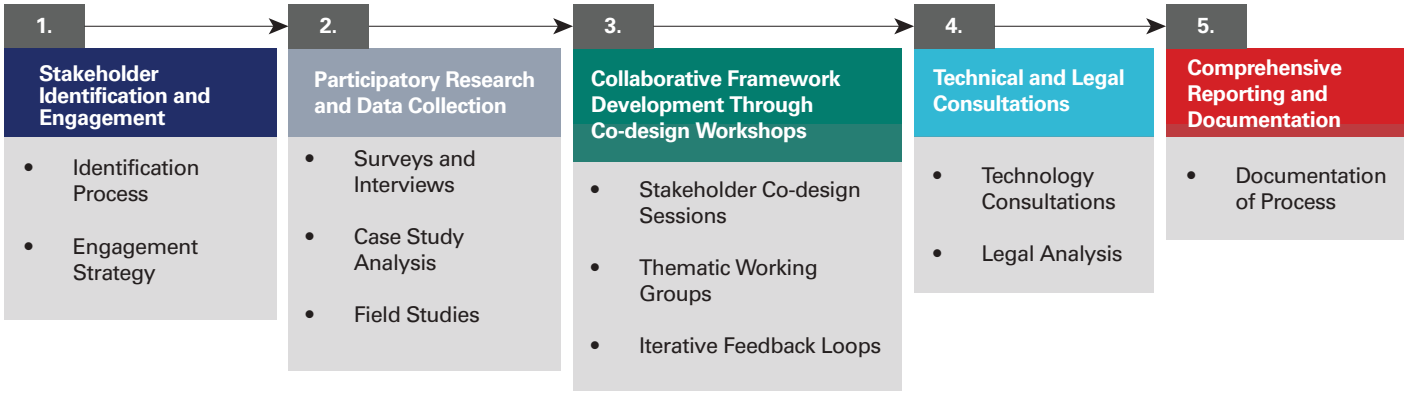


Figure 1: Methodology for Developing the Framework for Nigeria's Online Harms Protection Bill

Source: Advocacy for Policy and Innovation (API) Intelligence

This white paper outlines the development of a proposed framework for Nigeria's Online Harms Protection (OHP) Bill, underscoring a participatory approach that integrated contributions from various stakeholders. The methodology embraced collaborative methods, participatory research, and co-design, ensuring all stakeholders were actively involved at each stage to achieve a comprehensive, context-sensitive framework responsive to Nigeria's digital environment.

1. Stakeholder Identification and Engagement

Identification Process: At the outset, a thorough process identified key stakeholder groups, including government regulatory bodies (e.g., NITDA), digital rights organisations, technology companies, legal experts, content creators, and representatives from marginalised communities. Each group brought a unique perspective, ensuring the framework addressed varied interests and insights.

Engagement Strategy: Regular multi-stakeholder meetings and workshops were held to capture diverse perspectives. This collaborative effort was central to developing the framework, enabling iterative input and continuous feedback from groups directly impacted by online harms.

2. Participatory Research and Data Collection

Surveys and Interviews: Surveys and in-depth interviews were conducted to gather data on the prevalence and impact of online harms in Nigeria. Secondary data on victims of online abuse were analysed and provided valuable insights, helping to shape a framework that reflects Nigeria's specific digital challenges.

Case Study Analysis: The project team analysed various international frameworks, including Germany's NetzDG, the EU's Digital Services Act, and similar regulatory frameworks in South Africa and Kenya. Lessons learned from these examples informed the contextualisation of the Nigerian framework, helping to identify best practices and existing gaps in Nigeria's content moderation practices.

Field Studies: The team collaborated with its steering committee members to document the experiences of vulnerable groups, including children, women, and minority communities, in encountering online harms. This collaboration ensured the framework was sensitive to the needs of at-risk groups, providing tailored protection mechanisms in the final draft.

3. Collaborative Framework Development Through Co-design Workshops

Stakeholder Co-design Sessions: To ensure inclusivity, the team organised collaborative design sessions where stakeholders could co-create specific elements of the OHP framework. These workshops focused on critical areas such as intermediary liability, transparency standards, accountability mechanisms, and integrating a "duty-of-care" model. The contributions of each stakeholder group were synthesised to develop a framework that balances protection, rights, and regulatory requirements.

Thematic Working Groups: The development process included thematic groups from the steering committee dedicated to critical issues like data protection, content moderation, hate speech regulation, and child online safety. These groups provided recommendations that were integrated into the broader framework. This structure facilitated a deep dive into complex areas, promoting a more refined and practical approach to each issue.

Iterative Feedback Loops: Throughout the process, iterative reviews were conducted with thematic groups and external reviewers to refine the draft framework continuously. This allowed the team to respond to emerging issues and incorporate new insights, enhancing the framework's adaptability and robustness.

4. Technical and Legal Consultations

Technology Consultations: Engaging AI and data protection specialists was essential to developing the technical aspects of content moderation and algorithmic transparency. This collaboration ensured the framework's technical provisions were sound and feasible, addressing practical content monitoring and moderation challenges.

Legal Analysis: Legal experts rigorously evaluated the draft to ensure alignment with Nigeria's constitutional protections and international human rights standards. Emphasis was placed on balancing freedom of expression with the need to mitigate online harms, a central tenet in shaping an inclusive, rights-respecting framework.

5. Comprehensive Reporting and Documentation

Documentation of Process: Each stage was carefully documented, creating a transparent record of stakeholder contributions, workshop outcomes, and pilot results. This transparency strengthened the framework's legitimacy and provided a foundation for future adjustments.

The development of Nigeria's Online Harms Protection framework within this white paper was rooted in a participatory, inclusive methodology that prioritised stakeholder input and collaboration. This approach ensured the framework reflects diverse perspectives, respects digital rights, and provides actionable mechanisms for addressing online harms in Nigeria. The resulting OHP Bill framework embodies a balanced, adaptable, and context-sensitive model for digital safety, setting a standard for inclusive policy-making in Nigeria's digital landscape. The goal of this whitepaper is to create an aggregated regulatory framework to be promoted as a stakeholder led effort to develop Nigeria's online harms protection law.

Chapter 1

1.0 Introduction

In the dynamic digital age, where information access and connectivity thrive, the challenges to maintaining the integrity of online interactions are evident. Globally, digital content moderation (CM) strives to curb the dissemination of harmful materials, and Nigeria's approach reflects a dual strategy involving governmental oversight and self-regulation on online platforms.

Despite these global initiatives and the relative success of CM practices, the known inadequacies of CM practices are apparent and consequential. Self-regulatory measures often lack consistency and transparency, leading to accusations of bias and censorship. Automated systems need help with language and local nuances, resulting in over- or under-moderation. Human moderators face a devastating psychological toll whilst reviewing disturbing content. The global nature of the internet also introduces jurisdictional conflicts that demand a cooperative approach. A comprehensive policy framework encouraging collaboration and partnership between stakeholders is crucial to address these challenges effectively.

"Global experience indicates that content moderation (CM)-centered approaches, whether human-based or augmented by machines, have inherent limitations that necessitate several shifts in strategy. This acknowledges the need for evolving strategies to address online harms."

Recent studies indicate that nearly 90% of young adults globally have encountered harmful content. In Nigeria, over 50% of girls aged 15 to 25 have experienced online harassment or abuse[1]. Recognising the limitations of legislative initiatives and self-regulatory approaches globally, this white paper advocates a strategic shift from conventional CM to a proactive model of online harm protection. In proposing a new Nigerian regulatory framework that emphasises citizen protection, especially for vulnerable groups, this paper advocates a co-regulatory and "duty-of-care" model.

Nigeria's digital ecosystem, teeming with user-generated content, faces rising risks such as cyberbullying, child exploitation, and misinformation, particularly during critical periods like elections. For example, the Child Online Safety Index (COSI) highlights global cyber risks encompassing countries worldwide. Nigeria is particularly exposed to online threats that endanger children.

In Nigeria, regulatory efforts aim to curb harmful online content and ensure a safer digital environment. The Cybercrimes (Prohibition, Prevention) Act of 2015 criminalises various online offences, providing a legal basis for prosecuting illicit online activities and mandating internet service providers (ISPs) to manage content in line with the law. The National Information Technology Development Agency (NITDA) plays a crucial role in setting standards for CM through regulatory guidelines. *NITDA's Framework and Guidelines for Public Internet Access, released in 2019*, outlines service providers' responsibilities in ensuring the safe use of public Internet access by Nigerians and non-Nigerians alike. Nigeria has also seen the introduction of the Protection from Internet Falsehood and Manipulations Bill, known as the "Social

[1]Brain Builders Youth Development Initiative (BBYDI). "Factsheet Online Gender-Based Violence." BBYDI. September 2023. https://thebrainbuilders.org/wp-content/uploads/2024/05/Factsheet-Online-Gender-Based-Violence-4.pdf

Media Bill," which has been as controversial as it has been divisive. Despite concerns about its potential impact on freedom of speech, the bill seeks to regulate the spread of false information online and sanction disseminators of falsehood. Under the Internet Code of Practice, the Nigerian Communications Commission (NCC) actively regulates the exposure to objectionable, offensive, and potentially harmful content. It protects minors and vulnerable audiences online through its statutory and regulatory oversight of telecommunications service providers.

Even with these efforts considered, the urgency for a more robust online harm protection law is underscored by the persistent and evolving nature of online threats. Despite existing regulations, the complexities of the digital environment require a more comprehensive and collaborative approach to online safety for Nigerians.

Regulatory efforts on CM across Africa also reflect diverse approaches and peculiar challenges. Several nations have enacted legislation and guidelines to address harmful online content and protect digital users. Examples include *South Africa's Films and Publications Act, Kenya's Computer Misuse and Cybercrimes Act, Egypt's Cybercrime Law, Ethiopia's Hate Speech and Disinformation Prevention and Suppression Proclamation, Ghana's Electronic Communications Act,* and *Tanzania's Electronic and Postal Communications (Online Content) Regulations.* These illustrate the diverse regulatory measures across Africa to address CM challenges. From this, the continent's commitment to tackling online harms is evident. Yet, concerns arise about the potential impact on freedom of expression and the delicate balance between content regulation and digital rights. Legislative efforts globally, such as *Germany's NetzDG, Australia's Online Safety Act, the EU's DSA, Ireland's Media Regulation Act,* and the *UK's Online Safety Act,* reflect a commitment to online safety, mandating proactive measures against harmful content and cyber threats. These regulations have all faced significant challenges, primarily around the fear of potential censorship and the delicate balance between curbing harmful content and preserving free speech. Other concerns include the subjective nature of content interpretation, the risk of regulatory overreach, and the potential impact on innovation, especially for smaller online platforms. These legislative efforts present lessons and best practices for evolving climes like Nigeria and are the inspiration from which this white paper draws.

"This white paper proposes a shift from the current CM-centered approach toward a coherent, coordinated framework that guarantees citizens' rights while shielding society from the harms of the internet."

This white paper underscores the importance of regulatory measures broadly, and for children and minorities especially, and calls for a collaborative and data-driven approach to crafting a framework to ensure a safer digital environment for all. The proposed framework centres around a duty-of-care model built on accountability, transparency, and collaborative efforts.

In conclusion, this white paper acknowledges that while CM efforts are evolving in Nigeria, Africa, and globally, their intended impact to maximise the online community's well-being has been imperfectly achieved. The inadequacies of current CM practices highlight the need for a more coherent and practical framework that can adapt to the complexities and dynamics of the evolving digital landscape.

1.1 Definition of Terms: Conceptual Framework

This white paper outlines the development of a proposed framework for Nigeria's Online Harms Protection (OHP) Bill, underscoring a participatory approach that integrated contributions from various stakeholders. The methodology embraced collaborative methods, participatory research, and co-design, ensuring all stakeholders were actively involved at each stage to achieve a comprehensive, context-sensitive framework responsive to Nigeria's digital environment.

Concept	Definition
Harmful Content	Harmful online content is any material encountered online that can cause distress to an individual. This can vary widely and is often interpreted differently based on the subject's cultural, religious, and legal context[2]. The subjective nature of what constitutes harmful content means that what may be distressing to one individual may not be perceived as such by another. Online harms can manifest in various forms, including behaviours that cause physical or emotional injury. Such behaviours might include the sharing or sending of harmful information. Recognisable categories of harmful content include online abuse, cyberbullying, harassment, threats, impersonation, unsolicited sexual advances, violent imagery, content encouraging self-harm or suicide, and pornography. Additionally, harmful content may involve disseminating damaging information, such as misinformation and disinformation, generally called "fake news", which can have broader societal impacts beyond individual distress[4]. It can also include hate speech and the promotion of disturbing content such as drug use and other illicit activities[5]. Other forms of harmful content that have been identified include threats and intimidation, racism, indecent or abusive imagery, materials promoting terrorism or extremism, and various forms of cybercrime, including malware, scams, and phishing. Online child exploitation, defamation, incitement to commit crimes and slander are also recognised as harmful content that can have severe consequences for the well-being of individuals and the integrity of broader online communities[6].

[2]Keipi, Teo, et al. *"Online Hate and Harmful Content: Cross-National Perspectives."* Dec. 2016, www.researchgate.net/publication/311587458_Online_Hate_and_Harmful_Content_Cross-National_Perspectives, https://doi.org/10.4324/9781315628370.
[3]Woodhouse, John . "Research Briefing: Regulating Online Harms." Parliament.uk, House of Commons Library, 15 Mar. 2020, researchbriefings.files.parliament.uk/documents/CBP-8743/CBP-8743.pdf.
[4]Anderson, Janna, and Lee Rainie. "The Future of Truth and Misinformation Online." *Pew Research Center*, 19 Oct. 2017, www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/.
[5]"What Is Harmful Content Online?" *Digital Parenting Coach*, 6 Nov. 2023, www.digitalparentingcoach.com/blog/what-is-harmful-content. Accessed 3 Mar. 2024.
[6]Broadband Commission. "Child Online Safety: Minimising the Risk of Violence, Abuse and Exploitation Online." *Unesco.org*, Oct. 2019, unesdoc.unesco.org/ark:/48223/pf0000374365.

Concept	Definition
Hate Speech	The concept of hate speech is recognised internationally as a form of expression that can incite violence, discrimination, and hatred against individuals or groups based on specific characteristics. According to the *United Nations Strategy and Plan of Action on Hate Speech*, hate speech is defined as any kind of communication, whether verbal, written, or behavioural, that attacks or uses pejorative language with the intent to discriminate against a person or a group based on attributes such as their religion, ethnicity, nationality, race, colour, descent, gender, or other identity factors[7]. This definition underscores the potentially damaging impact of hate speech on social cohesion and individual dignity. It is not limited to spoken words but includes written materials, online content, and behaviours that convey hateful messages. The focus on the identity factors mentioned in the definition highlights the need for societies to protect vulnerable groups[8] such as racial and ethnic minorities, religious communities, refugees and immigrants, people living with disabilities, and women and girls, from speech that seeks to undermine their rights and existence. Efforts to combat hate speech often involve a combination of legal measures, public education, and policies promoting tolerance and diversity. However, addressing hate speech also involves navigating the thin line between protecting freedom of expression and preventing language that could lead to harm or discrimination. Many societies are actively working to address this complex terrain through various means, including legislation, community engagement, and international cooperation.–

[7]United Nations. UN Strategy and Plan of Action on Hate Speech. May 2019, www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf.
[8]These groups often face targeted hate speech that seeks to undermine their rights and existence, necessitating robust protective measures.

Concept	Definition
Protection from Online Harms	Protection from online harms refers to measures and strategies implemented to safeguard individuals, especially vulnerable groups such as children, from various types of damaging and dangerous content on the internet. This concept encompasses efforts to prevent exposure to cyberbullying, hate speech, extremist content, disinformation, and other forms of digital abuse that can lead to psychological, emotional, or even physical harm. In legal and policy contexts, the protection from online harms often translates into regulatory frameworks that impose duties on ISPs and digital platforms to monitor, report, and mitigate the presence of harmful content. The United Kingdom, for example, has been at the forefront of such legislative efforts with the introduction of the Online Safety Act of 2023, which aims to establish a statutory duty of care on digital service providers to protect users from harmful content(s)[9]. Similarly, the European Union's Digital Services Act (DSA) was designed to protect the digital space against the spread of illegal and harmful content and disinformation and to safeguard users' fundamental rights online. These legislative measures are part of a broader global movement to ensure digital platforms are more accountable for the content they host and to provide users with a safer online experience.

[9]UK Parliament. Online Safety Act 2023. Retrieved from https://bills.parliament.uk/bills/3137

Concept	Definition
Content Moderation (CM)	CM reviews and monitors (within predefined thresholds) user-generated content (UGC) online to ensure it meets specific standards and guidelines. It refers to the practice of controlling unwanted content in online spaces, whether that content is viewed as simply irrelevant (e.g. on an online forum with a specific topic), obscene, or illegal[10]. This includes removing inappropriate or offensive content and enforcing community guidelines and terms of service. When a user submits content to a website, that content is expected to undergo a screening process (known as the moderation process) to ensure that the content upholds the website's regulations and is not illegal, inappropriate, or harassing, amongst other criteria[11]. Thus, CM is the organised practice of screening UGC posted to internet sites, social media, and other online outlets to determine the appropriateness of the content for a given site, locality, or jurisdiction. The process can result in UGC being removed by a moderator acting as an agent of the platform or site. Increased Internet and social media platform penetration has led to increased UGC, creating a need for platforms and sites to enforce rules and relevant or applicable laws. This is because posting inappropriate content is considered a significant source of liability[12]. CM is particularly challenging as what constitutes CM may be contextual, difficult to define, often culturally subjective and legally ambiguous in some cases. This complexity is heightened in online environments where the information is partially or wholly derived from a large, diverse, and diffused user base[13]. There are three types of CM: human-based CM, automated CM, and a combination of both human and computerised mechanisms (i.e. the augmented model). Human moderation, also known as manual moderation, involves humans manually reviewing and monitoring UGC on an online platform to enforce platform-specific rules and guidelines. This helps protect online users by preventing unwanted, illegal, inappropriate content, scams, and harassment from appearing on the website. Automated CM, which relies on AI, automatically accepts, rejects, or sends UGCs based on the platform's rules and guidelines for human moderation. It is an efficient solution for online platforms aiming to ensure high-quality content goes live instantly while maintaining a safe user interaction environment[14]. Artificial Intelligence-based (AI-CM). AI-CM, often referred to as tailored AI moderation, involves the development of a machine learning model using data specific to an online platform to effectively and precisely identify undesirable UGCs. Through AI-CM, the system can automatically make highly accurate decisions regarding whether to reject, approve, or escalate content, thereby enhancing the efficiency of CM on the platform.

Concept	Definition
Protection from Online Harms and CM	Protection from online harms is a comprehensive approach encompassing various strategies and measures to safeguard users from potential dangers and negative online experiences. This concept involves not only the removal of illegal or inappropriate content but also the prevention of exposure to such content, as well as the promotion of digital literacy and the provision of support to individuals affected by online harms[15]. CM, on the other hand, is a subset of online harm protection. While CM is primarily reactive, focusing on dealing with harmful content after it is posted, protection from online harms is *proactive* and *reactive while protecting digital rights*[16]. Protection from online harms aims to create an environment where harmful content is less likely to be shared in the first place and where users are equipped with the knowledge and tools to protect themselves online. This broader approach includes legislative frameworks that set out a statutory duty of care for online companies to protect users from harmful content. In essence, CM is, and ought to be treated as a critical tool within the broader scope of online harm protection, which includes a more comprehensive range of policies, educational initiatives, and support mechanisms designed to foster a safer online ecosystem.

1.2 Online Harms: Categorisation and Legal Implications

"*Illegal content*" and *"harmful content"* are related concepts but differ in their legal and practical implications:

- **Illegal Content:**
  - Definition

    Illegal content violates established laws, regulations, or statutes within a specific jurisdiction. This includes content contravening criminal, civil, intellectual property rights, or regulatory provisions.[17]

  - Examples

    Illegal content encompasses a wide range of material, including but not limited to copyrighted material distributed without authorisation, pirated software, child exploitation material, terrorist propaganda, hate speech, defamation, fraud, and incitement to violence.

  - Legal Consequences

    Producing, distributing, or possessing illegal content can result in legal sanctions, including fines, imprisonment, and civil liability. Law enforcement agencies enforce laws related to illegal content, and individuals or entities found guilty may face criminal prosecution or civil lawsuits.[18]

- **Harmful Content:**
  - Definition

    Harmful content refers to material that has the potential to cause harm to individuals, groups, or society, even if it does not necessarily violate specific laws. Harm can manifest in various forms, including physical harm, psychological distress, emotional harm, reputational damage, or societal harm.[19]

  - Examples

    Harmful content includes content that promotes violence, hate speech, discrimination, harassment, bullying, disinformation, misinformation (in some cases), graphic or explicit material, and content that glorifies harmful behaviours.

  - Practical Considerations

    While harmful content may not always be explicitly illegal, it can still harm individuals and communities. Content moderation policies and community guidelines established by online platforms often prohibit harmful content from maintaining a safe and respectful online environment. However, the boundaries between harmful and permissible content can be subjective and context-dependent, leading to debates and challenges around content moderation decisions.[20]

In summary, illegal content refers to material that violates established laws and regulations, while harmful content encompasses material that has the potential to cause harm, regardless of its legality. While there may be overlaps between the two categories, not all harmful content is necessarily illegal, and vice versa.

[17]Herbert Smith Freehills. "What is 'illegal content' and what are the key duties under the Online Safety Act?" October 2024. *herbertsmithfreehills.com.* https://www.herbertsmithfreehills.com/insights/2024-10/what-is-illegal-content-and-what-are-the-key-duties-under-the-osa#:~:text=The%20term%20'illegal%20content'%20is,constitutes%20a%20'relevant%20offence'.

[18]Crane, Amy. "What Are the Penalties for Illegally Downloading Content?" Super Lawyers. February, 2024. https://www.superlawyers.com/resources/intellectual-property/what-are-the-penalties-for-illegally-downloading-content/

[19]Ofcom. "Illegal and Harmful Content." 2024. https://www.ofcom.org.uk/

[20]Vogelezang Francesco. "Illegal vs Harmful Online Content." Internet Just Society. December, 2020. https://www.internetjustsociety.org/illegal-vs-harmful-online-content

[21]Herbert Smith Freehills. "What is 'illegal content' and what are the key duties under the Online Safety Act?" October 2024. *herbertsmithfreehills.com.* https://www.herbertsmithfreehills.com/insights/2024-10/what-is-illegal-content-and-what-are-the-key-duties-under-the-osa#:~:text=The%20term%20'illegal%20content'%20is,constitutes%20a%20'relevant%20offence'.

1.2.2 Legal but Harmful

These acts, while potentially causing harm, may not be explicitly illegal in a given jurisdiction. They raise ethical concerns and can have negative consequences, even if not criminal.[21] Examples include politically divisive ads that may comply with regulations but exacerbate societal tensions.

Germany: The NetzDG Approach

The Network Enforcement Law (NetzDG)[22] requires social networks with more than 2 million registered users in Germany to exercise a local takedown of obviously illegal content (e.g. a video or a comment) within 24 hours after notification. Where the illegality is not apparent, the provider typically has up to seven days to decide on the case.

On an exceptional basis, it can take longer and be referred to a joint industry body accredited as an institution of regulated self-regulation.

To qualify for removal under NetzDG, content must fall under one of the 21 criminal statutes in the German Criminal Code (StGB). Online platforms also evaluate content under their global community guidelines, and the content is removed if it violates these global guidelines. If the content does not fall under these policies but is identified as illegal according to one of the 21 statutes of the StGB to which NetzDG refers (§ 1 III NetzDG) or any other local law, the removal of the content is restricted locally. The NetzDG also requires social networks to create and publish a bi-annual report about the handling of such complaints (transparency report).[23]

Criminal offences provided for under local laws, which are referred to under the NetzDG, include:

Hate Speech or Political Extremism

- Incitement to hatred.[24]
- Defamation of religions, religious and ideological associations.

Terrorist or Unconstitutional Content

- Dissemination of propaganda material of unconstitutional organisations.[25]
- Using symbols of unconstitutional organisations.

[22]The German Bundestag. "Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act)." July 2017, www.bmj.de/SharedDocs/Downloads/DE/Gesetzgebung/RefE/NetzDG_engl.pdf?__blob=publicationFile&v=4.
[23]Library of Congress."Germany: Network Enforcement Act Amended to Better Fight Online Hate Speech." loc.gov. https://loc.gov/item/global-legal-monitor/2021-07-06/germany-network-enforcement-act-amended-to-better-fight-online-hate-speech/
[24]Spittlegerber and Wilde-Detmering. "Germany's New Hate Speech Act in Force: What Social Network Providers Need to do Now." Technology Law Dispatch. October 2017. https://www.technologylawdispatch.com/2017/10/social-mobile-analytics-cloud-smac/germanys-new-hate-speech-act-in-force-what-social-network-providers-need-to-do-now/
[25]Gov.UK Report online material promoting terrorism or extremism. https://www.gov.uk/report-terrorism#:~:text=articles%2C%20images%2C%20speeches%20or%20videos,videos%20of%20terrorist%20attacks.

Violence

- Dissemination of depictions of violence.[26]

Harmful or Dangerous Acts

- Public incitement to crime.
- Breach of the public peace by threatening to commit offences.
- Defamation or insult.

Privacy

- Violation of intimate privacy by taking photographs.[28]

Sexual Content

- Distribution, acquisition and possession of child pornography.[29]
- Distribution of pornographic performances by broadcasting, media services, or telecommunications services.

Misinformation and Disinformation

- Spreading false or misleading information, often intentionally, to deceive or manipulate others.[30]

Doxxing

- Publicly revealing private or identifying information about someone, often with malicious intent.[31]

Non-consensual Intimate Sharing

- Sharing private or intimate images or videos of someone without their consent.[32]

Hate Speech Bordering on but not Explicitly Illegal

- Language that is offensive, hateful, or discriminatory but may not meet the legal threshold for hate speech in specific jurisdictions.

Online Harassment

- Engaging in behaviour that annoys, bothers, or alarms someone online without necessarily meeting the criteria for cyberbullying.[33]

---

[26]Busching Robert et al. Violent Media Content and Effects. *Oxford Research Encyclopedia, CommunicationPublisher: New York: Oxford University Press.* March 2016. https://www.researchgate.net/publication/323784251_Violent_Media_Content_and_Effects

[27]Gillet Mathew. Proving Online Incitement of International Crimes: Expert Evidence in the Digital *Era. Essex Law Research.* February 2024. https://essexlawresearch.uk/2024/02/13/proving-online-incitement-of-international-crimes-expert-evidence-in-the-digital-era/

[28]Ghazinour Kambiz and Ponchak John. Hidden Privacy Risks in Sharing Pictures on Social Media. *Procedia Computer Science* Volume 113, 2017, Pages 267-272. 2017. https://www.sciencedirect.com/science/article/pii/S1877050917317775

[29]E Safety Commissioner. 'What is Illegal and Restricted Online Content?' https://www.esafety.gov.au/report/what-is-illegal-restricted-content

[30]American Psychological Association. 'Misinformation and disinformation. apa.org.' https://www.apa.org/topics/journalism-facts/misinformation-disinformation

[31]Kaspersky Resource Centre. "What is Doxxing – Definition and Explanation" https://www.kaspersky.com/resource-center/definitions/what-is-doxing

[32]Cyberbullying and the Non-consensual Distribution of Intimate Images. https://www.justice.gc.ca/eng/rp-pr/other-autre/cndii-cdncii/p6.html

[33]What is Online Harassment Durham University. https://reportandsupport.durham.ac.uk/support/what-is-online-harassment#:~:text=Definition,humiliation%20in%20an%20online%20setting.

1.2.3 Other Online Harms

This category encompasses harmful online activities that may not be illegal or directly cause immediate harm but can have detrimental societal and individual consequences:

Addiction to Online Platforms and Activities:
Excessive use of social media or online games leads to negative impacts on mental health, relationships, and productivity.[34]

Echo Chambers and Filter Bubbles:
Exposure to information that reinforces existing beliefs and limited exposure to diverse viewpoints, potentially leading to polarisation and hindering critical thinking.[35]

Privacy Concerns:
Unlawful data collection, profiling, and targeted advertising based on personal information.[36]

It is important to note that the legal classification of online harms can vary depending on specific jurisdictions and evolving legal interpretations. Additionally, the lines between these categories can be blurry; for example, hate speech that initially falls under the "harmful, yet legal" category can become criminal if it incites violence or specific threats.

Therefore, addressing online harms requires a multifaceted approach, including legal frameworks, ethical considerations within technology development and user behaviour, and individual awareness and critical thinking skills to navigate the online world responsibly.

[34]Cash et al. 'Internet addiction: A brief summary of research and practice', *Current Psychiatry Reviews*, 8(4), pp. 292-298. November, 2012. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2719452/

[35]Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W. and Starnini, M. (2021) 'The Echo Chamber Effect on Social Media', Proceedings of the National Academy of Sciences, 118(9), e2023301118. February 2021. https://www.pnas.org/content/118/9/e2023301118

[36]Ban Liyuan. "Data privacy and protection in communication networks." Applied and Computational Engineering 38(1):131-138. DOI:10.54254/2755-2721/38/20230541. February 2024. https://www.researchgate.net/publication/378435413_Data_privacy_and_protection_in_communication_networks

1.3 Potential Impact of Algorithms

This section of the whitepaper delves into the current use of AI in content moderation, the increasing reliance on algorithmic solutions triggered by regulatory pressures and scalability challenges, and the resultant harms posed by these algorithms. It concludes with recommendations for policy interventions.

The digital age has ushered in unprecedented interconnectivity and information flow facilitated by social media platforms and various online services. Artificial Intelligence (AI) has become integral in managing these vast data streams, particularly online harm moderation. AI algorithms have increasingly been used to maintain online environments, ensuring they remain conducive to positive user experiences by filtering out harmful content.[37] There have been multiple use cases of deployment of AI for content moderation.[38]

AI's role in content moderation has expanded from simply replicating human moderation decisions to proactive content monitoring.[39] This shift is primarily due to the challenge of detecting augmented content and the sheer volume of user-generated content, which renders manual moderation impractical and inefficient.[40] Also, the advent of generative AI poses new challenges, with malicious actors using deepfakes, voice clones, and synthetic media to propagate misleading narratives.[41] AI algorithms now play a pivotal role in identifying and mitigating various forms of harmful content, including hate speech, misinformation, and explicit material, safeguarding user well-being, and upholding community standards.[42]

1.3.1 Use of AI in Content Moderation

AI is being deployed to facilitate the rapid and efficient scanning of vast amounts of online content and to identify content that is against community standards, and that may violate local law.[43] This capability is precious in detecting and mitigating hate speech, fake news, and explicit materials. Traditional manual moderation methods, while necessary, are increasingly supplemented by AI to address the limitations of scalability and speed. Furthermore, AI moderation tools have been touted to help reduce the exposure of human moderators to psychologically harmful content, thus preserving their mental wellbeing.[44]

The growing obligations on platforms to maintain a responsible digital environment and the impracticality of scaling human moderation to match the volume of user-generated content have incentivised the shift towards automated, AI-driven moderation.[45] Legal attractiveness, cost-efficiency, and scalability make AI appealing for platforms, which have claimed that it helps them comply with regulatory demands without compromising moderation quality or speed.[46]

---

[37] "The Future of AI in Content Moderation and Censorship." Faster Capital, fastercapital.com/topics/thefuture-of-ai-in-content-moderation-and-censorship.html. Accessed 2 Mar. 2024.

[38] Kniazieva, Yuliia. "AI Content Moderation for Responsible Social Media Practices." Labelyourdata.com, 30 Mar. 2023, labelyourdata.com/articles/ai-content-moderation.

[39] "Role of AI in Content Moderation and Censorship." Faster Capital, https://fastercapital.com/content/Roleof-ai-in-content-moderation-and-censorship.html, Accessed 2 March 2024

[40] Francisco. "Why Moderation Has Become Essential for UGC." Checkstep, 10 Jan. 2024, www.checkstep.com/why-moderation-has-become-essential-for-ugc/. Accessed 02 March 2024.

[41] Atleson, Michael. "Chatbots, Deepfakes, and Voice Clones: AI Deception for Sale." Federal Trade Commission, 20 Mar. 2023, www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clonesai-deception-sale. Accessed 2 March 2024.

[42] Ibid. n3.

[43] Somers, Charlotte. "Ensuring Online Safety - the Role of Artificial Intelligence in Combatting Illegal Content Online." www.law.kuleuven.be, 27 June 2023, www.law.kuleuven.be/ai-summerschool/blogpost/Blogposts/AI-combatting-illegal-content-online. Accessed 2 March 2024.

[44] Newton, Casey. "The Secret Lives of Facebook Moderators in America." The Verge, 25 Feb. 2019, www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-traumaworking-conditions-arizona. Accessed 2 Mar. 2024.

[45] Lardin, Juliette. "The Importance of Scalability in AI Content Moderation." Checkstep, 31 Dec. 2023, www.checkstep.com/the-importance-of-scalability-in-ai-content-moderation/. Accessed 2 Mar. 2024.

[46] Griffin, Rachel. "Algorithmic Content Moderation Brings New Opportunities and Risks." Centre for International Governance Innovation, 23 Oct. 2023, www.cigionline.org/articles/algorithmic-contentmoderation-brings-new-opportunities-and-risks/. Accessed 2 Mar. 2024.

1.3.2 Limitations of Human Moderation

Traditional content moderation methods, which rely heavily on human moderators to review and filter content, face significant challenges in the digital era.The volume of usergenerated content on major platforms is staggering, with millions of posts, comments, images, and videos uploaded every minute. Human moderators cannot feasibly review this deluge of content in real-time, leading to delays in removing harmful content.[47] This latency allows harmful material to remain accessible and downloadable, potentially causing distress or harm to an exponential number of users. Moreover, the manual review process is slow and labour-intensive and exposes moderators to distressing content, risking their mental health.

Additionally, human judgement is inherently subjective; different moderators may interpret content standards differently, leading to inconsistent content enforcement. This variability can undermine user trust in a platform's moderation policies.[48]

1.3.3 Algorithmic Harms and Impact

The potential for AI algorithms to inadvertently perpetuate harm is a significant concern. Biased decision-making, stemming from flawed training data, can reinforce stereotypes and marginalise communities.[49] For instance, algorithms trained on data from one ethnic group may misinterpret or unfairly target content from other ethnicities, leading to racial bias. Similarly, gender bias arises when algorithms trained on datasets with a predominance of male voices do not recognise or correctly interpret content from women.[27] This silences voices and perpetuates a cycle of exclusion and bias in digital spaces.

The use of generative AI by malicious actors to create deepfakes and synthetic media introduces a new dimension of risk.[51] These technologies can manufacture highly convincing yet entirely false content, from fake news to counterfeit audiovisual materials, further complicating distinguishing between legitimate and harmful content. An example includes the creation of politically motivated deepfakes aimed at manipulating elections or inciting social unrest. In February 2024, a video depicting a deceased former Indonesian president endorsing a political party in a recent election raised significant concerns.[52] This underlies the dual-edged nature of AI advancements.

The harms associated with algorithmic content moderation extend beyond direct bias and discrimination. For instance, the echo chamber effect, where algorithms curate content that reinforces a user's existing beliefs, can exacerbate social divisions and polarisation.[53] Similarly, the overreliance on algorithms can lead to the suppression of free speech, where legitimate content is mistakenly flagged and removed, stifling public discourse.[54] This, in turn, spreads harmful content and radicalises viewers.

Social media platforms, driven by the business model of attracting advertisers and strengthening revenue streams, maximise user engagement, a critical metric that influences their success and the retention of users.[55] This incentive structure can lead to the proliferation of sensationalist, extremist, or polarising content, as such material is more likely to generate clicks, shares, and prolonged engagement.

---

[47] Rizoiu, Marian-Andrei, and Philipp Schneider. "Can Human Moderators Ever Really Rein in Harmful Online Content? New Research Says Yes." The Conversation, 14 Aug. 2023, theconversation.com/can-humanmoderators-ever-really-rein-in-harmful-online-content-new-research-says-yes-209882. Accessed 2 Mar. 2024.

[48] bid. n16

[49] *Bias in Algorithms – Artificial Intelligence and Discrimination*. European Union Agency for Fundamental Rights, 2022.Pg 69 https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf

[50] Castillo, Dianne. "The Gender Data Gap in AI: Confronting Bias in Machine Learning." Seldon, 28 Feb. 2023, www.seldon.io/the-gender-data-gap-in-ai. Accessed 2 Mar. 2024.

[51] Ferrara, Emilio. "GenAI against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models." Journal of Computational Social Science, 22 Feb. 2024, https://doi.org/10.1007/s42001-024-00250-1. Accessed 2 March 2024.

[52] Chen, Heather. "AI "Resurrects" Long Dead Dictator in Murky New Era of Deepfake Electioneering." CNN, 12 Feb. 2024, www.cnn.com/2024/02/12/asia/suharto-deepfake-ai-scam-indonesia-election-hnk-intl/index.html. Accessed 2 Mar. 2024.

[53] "Ethical Considerations of Content." Faster Capital, https://fastercapital.com/startup-topic/Ethical-Considerations-of-Content.html, Accessed 2 March 2024

[54] Solidity Law. "The Role of AI in Content Moderation: Free Speech, Censorship, and Legal Liability." www.linkedin.com, 7 July 2023, www.linkedin.com/pulse/role-ai-content-moderation-free-speech-censorshiplegal-liability. Accessed 2 Mar. 2024.

[55] Kapp, Brandon. "Profit-Driven Echo Chambers: Unveiling the Illusion of Diverse Beliefs on Social Media." www.linkedin.com, 14 June 2023, www.linkedin.com/pulse/profit-driven-echo-chambers-unveiling-illusiondiverse-brandon-kapp. Accessed 2 Mar. 2024.

The prioritisation of engagement over well-being has broader societal implications. Despite the best efforts of platforms, algorithms promote content that reinforces users' beliefs, creating filter bubbles and echo chambers.[56] These phenomena can significantly affect social cohesion as individuals become more entrenched in their viewpoints, less tolerant of opposing perspectives, and more susceptible to misinformation.

AI algorithms, trained on datasets that lack diversity, often need to recognise the nuances of speech, culture, and expression of marginalised communities. This can lead to ethnic bias, where content from specific groups is wrongly flagged or suppressed while the same expressions from dominant groups pass through unchecked. For example, algorithms trained on data from specific ethnic groups may inaccurately moderate content from other ethnic backgrounds due to misunderstood context or slang, leading to digital exclusion.

Determining what constitutes harmful content is inherently subjective, with significant variations across cultures, legal systems, and individual perceptions.[57] Content considered harmful or extremist in one context might be seen as a legitimate exercise of free speech in another. This subjectivity complicates the task of programming algorithms to accurately identify harmful content, often leading to over-moderation or under-moderation, which carries significant consequences for public discourse and democratic engagement.[58]

False positives, where benign content is mistakenly flagged or removed, can suppress free speech and limit the diversity of online discourse.[59] Conversely, false negatives, where harmful content remains undetected, can allow damaging narratives to increase, causing real-world harm. The evolving nature of online speech and the myriad forms of harmful content make this an ongoing and complex endeavour.

Outsourced software and complex AI supply chains need to be more transparent about the lines of accountability. When harmful or biased moderation occurs along the line, tracing the source of the questionable decision-making process is challenging, complicating efforts to rectify issues or hold entities accountable.[60]

[56] Ibid.

[57] Akdeniz, Yaman. "Freedom of Expression on the Internet: A Study of Legal Provisions and Practices Related to Freedom of Expression, the Free Flow of Information and Media Pluralism on the Internet in OSCE Participating States." Pg 19. Organisation for Security and Cooperation in Europe (OSCE), 2012. https://www.osce.org/files/f/documents/c/9/105522.pdf

[58] Sartor, Giovanni, and Andrea Loreggia. "The Impact of Algorithms for Online Content Filtering or Moderation "Upload Filters." Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies PE, European Parliament, Sept. 2020. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf Pg 47

[59] Cambridge Consultants. "Use of AI in Online Content Moderation 2019 Report (Produced on Behalf of OFCOM)." Pg 37. OFCOM, 2019. https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridgeconsultants-ai-content-moderation.pdf

[60] Brown, Ian. "Expert Explainer: Allocating Accountability in AI Supply Chains." Ada Lovelace Institute, 29 June 2023. https://www.adalovelaceinstitute.org/resource/ai-supply-chains/. Accessed 2 March 2024.

1.3.4 Limitations of Algorithm Moderation

Despite its benefits, AI algorithms have drawbacks, mainly when used in CM. They can inadvertently perpetuate harm through biased decision-making, the reinforcing of stereotypes, and failure to interpret nuanced content accurately.[61] Biases in training data can lead to discriminatory outcomes, while the lack of ethical or risk management frameworks amplifies harmful narratives.[62] Misrepresentations and the amplification of harmful content can deepen social divides, erode trust in digital platforms, and undermine the integrity of public discourse.[63] The harms associated with algorithmic moderation extend beyond individual bias or discrimination, threatening social cohesion and democratic processes.

Lastly, the challenges in accurately identifying and moderating harmful content can lead to over-censorship or the unchecked spread of damaging narratives.[64]

1.3.5 Recommendations

- Regulating AI-CM, created to regulate harmful content, presents a new set of complex legal and ethical challenges. There is a delicate balance between addressing harmful content and safeguarding freedom of expression. Overly stringent regulations may incentivise platforms to adopt conservative content removal policies, potentially stifling legitimate speech. Conversely, lax regulations might not adequately protect users from harm.

To mitigate these risks and harness the full potential of AI in content moderation, a multifaceted approach is necessary:

- To minimise biases, algorithms should be trained on diverse datasets that accurately reflect the local user base. This can be achieved by incorporating a wide range of cultural, linguistic, and demographic data, like initiatives undertaken by major tech companies to enhance speech recognition technologies across diverse languages and dialects. Ensuring the representativeness of training data can significantly reduce biases and improve the accuracy of content moderation across different communities.

[61] Newstead, Toby, et al. "How AI Can Perpetuate – or Help Mitigate – Gender Bias in Leadership." Organizational Dynamics, vol. 52, no. 4, 7 Sept. 2023,

[62] Ferrara, Emilio. "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies." Sci, vol. 6, no. 1, 26 Dec. 2023, pp. 3–3, www.mdpi.com/2413-4155/6/1/3, https://doi.org/10.3390/sci6010003. Accessed 2 March 2024.

[63] Bereskin, Cassidy. "Parliamentary Handbook on Disinformation, AI and Synthetic Media by the Parliamentarian." Commonwealth Parliamentary Association (CPA), 2023, issuu.com/theparliamentarian/docs/cpa_handbook_ai_disinformation_synthetic_media_onl. Accessed 2 Mar. 2024.

[64] United Nations. "Moderating Online Content: Fighting Harm or Silencing Dissent?" OHCHR, 23 July 2021, www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent. Accessed 2 Mar. 2024.

- While AI can screen content at scale, human moderators are essential for nuanced decision-making, especially in complex or sensitive contexts. Policies should enforce a constructive collaboration between AI capabilities and human judgement. Incorporating a human-centric approach in AI moderation, where human moderators work alongside AI to review content, ensures a more nuanced understanding of context and reduces the risk of errors.
- Algorithms should undergo periodic reviews to assess their impact on different demographics and refine their accuracy in identifying harmful content. This should be a legal obligation. Regular evaluation and auditing of algorithms help ensure that AI systems do not inadvertently perpetuate biases or facilitate harm. These audits should involve third-party assessments, providing an independent review of algorithmic performance and impact on various user groups.

- CM algorithms must be designed reasonably, ensuring fairness. Ensuring fairness requires that algorithms do not disproportionately silence or harm marginalised communities.[65] These algorithms also require accountability, incorporating mechanisms for oversight and user appeals process against moderation decisions, alongside platform responsibility for addressing wrongful content actions. Furthermore, transparency is crucial, necessitating that these platforms be transparent about the workings of their algorithms, the basis for content decisions, and efforts to counteract bias, thereby building trust and allowing for external review. Algorithms should be designed to enable the scrutiny of their decisionmaking processes to identify and correct bias.

- Tackling the intricacies of online harms necessitates a collaborative effort across various sectors. Policy intervention should establish clear guidelines for online harms that balance reducing harm and safeguarding freedom of expression while supporting the creation of industry standards and promoting research into ethical content moderation practices. Platforms should be encouraged to collaborate with external experts, exchange best practices, enhance moderation technologies through research, and develop advertising and content promotion algorithms that prioritise user well-being over engagement. Academics play a crucial role in assessing the effects of algorithmic moderation, innovating new ways to detect harmful content, and gauging the success of different moderation strategies. Additionally, civil society, including user groups, advocacy organisations, and impacted communities, should actively participate in formulating content moderation policies to ensure they uphold inclusivity and respect for human rights.

AI algorithms represent a significant advancement in content moderation, offering scalable solutions to the challenges of the digital age. However, a balanced approach incorporating ethical considerations, human oversight, and regulatory compliance is essential to fully realising AI's benefits whilst mitigating inherent risks. By adopting the recommended interventions, policymakers can ensure that AI serves as a force for good in the ongoing effort to maintain safe and inclusive online environments.

[65]Yaghi, Husam. "The Dark Side of Algorithms." *www.linkedin.com,* 24 Nov. 2023, www.linkedin.com/pulse/dark-side-algorithms-husam-yaghi-ph-d-ppu1e. Accessed 3 Mar. 2024.

1.4 *Duty-of-Care*

Disinformation, hate speech, and political polarisation are evident problems caused by the growing relevance of ICT in contemporary societies. To address these issues, decisionmakers and regulators worldwide continue to discuss the role of digital platforms in CM and in curtailing harmful content produced by third parties.

However, intermediary liability rules require a balance that avoids the risks arising from the circulation, at scale, of harmful content and the risk of censorship if excessive burdens force content providers to adopt a risk-averse posture in content moderation. This white paper examines the trend of altering intermediary liability models to include "*duty-of-care*" provisions, describing three models in Europe, North America, and South America.

Under Section 230 of the *Communications Decency Act,* the American model grants broad immunity to platforms for third-party content and content moderation. The previous European model under the E-Commerce Directive provided a "notice and takedown" approach, allowing platforms immunity with conditions. The Brazilian *Internet Bill of Rights* model grants immunity but will enable courts to order content removal.

These models have evolved to incorporate a "*duty-of-care*" approach, placing better monitoring and takedown obligations on platforms. For example, Germany's NetzDG law requires quick removal of "criminally punishable" content and improved transparency. The *EU's Digital Services Act* imposes "due diligence obligations" on platforms as a duty of care. The proposed Brazilian "*Fake News Bill* " focuses on platform transparency and user rights around content moderation.

In this white paper, we adopt a definition of "duty of care" as the legal obligation placed on internet service providers, social media platforms, search engines, and other online intermediaries to take reasonable measures to avoid harm to users from content transmitted or stored on their platforms. This represents a shift from the previous model of minimising interference in online content.

Duty-of-care models aim to balance limiting harmful content with protecting freedom of expression. The emerging duty-of-care approach represents a significant shift in intermediary liability, moving platforms from a "dumb pipe" model towards a more intelligent and dynamic model, with greater responsibility for moderating user-generated content and addressing its associated risks. We propose carefully considering these evolving content moderation frameworks' effectiveness and human rights implications with a duty-of-care-centred focus.

[68]Keller, Daphne. 'Systemic duties of care and intermediary liability.' *Stanford Center for Internet and Society.* May 2020. https://cyberlaw.stanford.edu/blog/2020/05/systemic-duties-care-and-intermediary-liability

1.5 Intermediary Liability

Intermediary liability refers to the legal responsibility of intermediaries such as internet service providers (ISPs), social media platforms, search engines, web hosting companies, and content delivery networks for the content transmitted or stored on their platforms.[67]

In some jurisdictions, platforms are required to act If the content is illegal or infringes on the rights of others. Increasingly, these intermediaries can be held liable in some jurisdictions where it may be construed that platforms have failed to act where facts establish that platforms had all the information required to act to prevent harm.

1.6 A Co-regulatory Approach

A co-regulatory approach to online harms protection involves collaborative efforts between governments, stakeholders (such as civil society) and internet platforms to establish a balanced framework for addressing harmful content. This approach acknowledges the shared responsibility of both parties to ensure a safer online environment while respecting freedom of expression. Under a co-regulatory model, governments set overarching policy objectives, legal requirements, and oversight mechanisms to guide content moderation practices. This includes defining standards for identifying and removing harmful content, promoting transparency, and safeguarding users' rights.

Internet platforms actively participate in developing and implementing content moderation practices to meet regulatory standards. They utilise their expertise and resources to enforce these standards effectively while maintaining the integrity of their platforms.

Ongoing dialogue and collaboration between governments and platforms are essential to a co-regulatory approach. This includes regular communication to refine content moderation strategies, address emerging challenges, and ensure alignment with regulatory objectives.

Key features of a co-regulatory framework include mechanisms for transparency, user appeals, and balancing against competing rights such as freedom of expression. These mechanisms help to foster accountability, trust, and legitimacy in the content moderation process.

Overall, a co-regulatory approach leverages the strengths of both the public and private sectors to tackle the complex and evolving issues of online content moderation and online harms protection, ultimately promoting a safer and more inclusive digital environment.

[67] "Intermediary Liability & Content Regulation." Global Network Initiative, globalnetworkinitiative.org/what-we-do/empower-policy/intermediary-liability-content-regulation/#:~:text=%E2%80%9CIntermediary%20liability%E2%80%9D%20describes%20the%20allocation,kinds%20for%20regulated%20content%20categories.

Chapter 2

2.0 Nigeria's Online Harm Landscape

Nigeria has over 103 million internet users, one-fourth of whom have social media access.[68] Approximately 14% of the country's population of about 220 million people[69] are social media users. The types of online harms these users may be susceptible to include all harms related to the production, distribution and consumption of online content.[70] Clearly defining these harms may be challenging as many of what constitutes online harms based on this categorisation may be contextual and cross-cutting. However, against the backdrop of recent events like the COVID-19 pandemic and the 2023 Nigerian elections, issues such as misinformation and disinformation *("fake news")* have been on the front burner of the subject of online harms in Nigeria.[71] These issues, in addition to other online content safety threats, including exposure to harmful, violent and illegal content, cyberbullying, as well as the challenge of online child exploitation, have energised conversations and legislative efforts to provide a legally binding framework to protect Nigerians.

The Nigerian digital landscape is saturated with a diverse array of UGC. Unfortunately, there is also a prevalence of illegal and harmful material, including hate speech, misinformation, disinformation, cyberbullying, online child pornography, revenge porn, harassment, threats, gender-based violence, and terrorism.[72]

"A fundamental duty of a state is the preservation of the rights of its citizens, including digital rights and the protection of these citizens from all categories of harmful incidents…… This emphasizes the state's responsibility in ensuring digital safety"

[68]Kemp, Simon. "Digital 2023: Nigeria." DataReportal – Global Digital Insights, 23 Feb. 2024, https://datareportal.com/reports/digital-2024-nigeria.
[69]UNFPA - United Nations Population Fund. "World Population Dashboard-Nigeria." www.unfpa.org. www.unfpa.org/data/world-population/NG.
[70]Grant, Julie, et al. Toolkit for Digital Safety Design Interventions and Innovations: *Typology of Online Harms-Insight Report.* World Economic Forum, Aug. 2023.
[71]Egwu Patrick. '"We can't do this alone": Nigerian fact-checkers teamed up to debunk politicians' false claims at this year's election', *Reuters Institute.* September 2023. https://reutersinstitute.politics.ox.ac.uk/news/we-cant-do-alone-nigerian-fact-checkers-teamed-debunk-politicians-false-claims-years-election
[72]Oyeniyi Joshua. 'Clicks, clout, and chaos: Content and cybercrime in Nigeria.' *Punch Newspaper.* October 2024. https://punchng.com/clicks-clout-and-chaos-content-and-cybercrime-in-nigeria/

Figure 2 below reflects case examples that shed light on the dangers and consequences of the various ways harmful content online manifests. The 2023 Nigerian general election exemplifies the pervasive threats of misinformation and disinformation, significantly impacting the election's integrity and credibility. Allegations of domestic interference and coordinated inauthentic behaviours involving politicians paying social media influencers to create fake accounts or use their authentic presence online to spread false narratives, sharing misleading information, and targeting specific individuals have led to the erosion of public trust, manipulation of voter behaviour, and potential compromise of election fairness.[73] Perpetrators routinely utilise online platforms, exacerbating challenges for electoral institutions.

Case Examples of Threats and Impacts of Online Harms in Nigeria

Figure 2: Case Examples of Threats and Impact of Online Harm in Nigeria

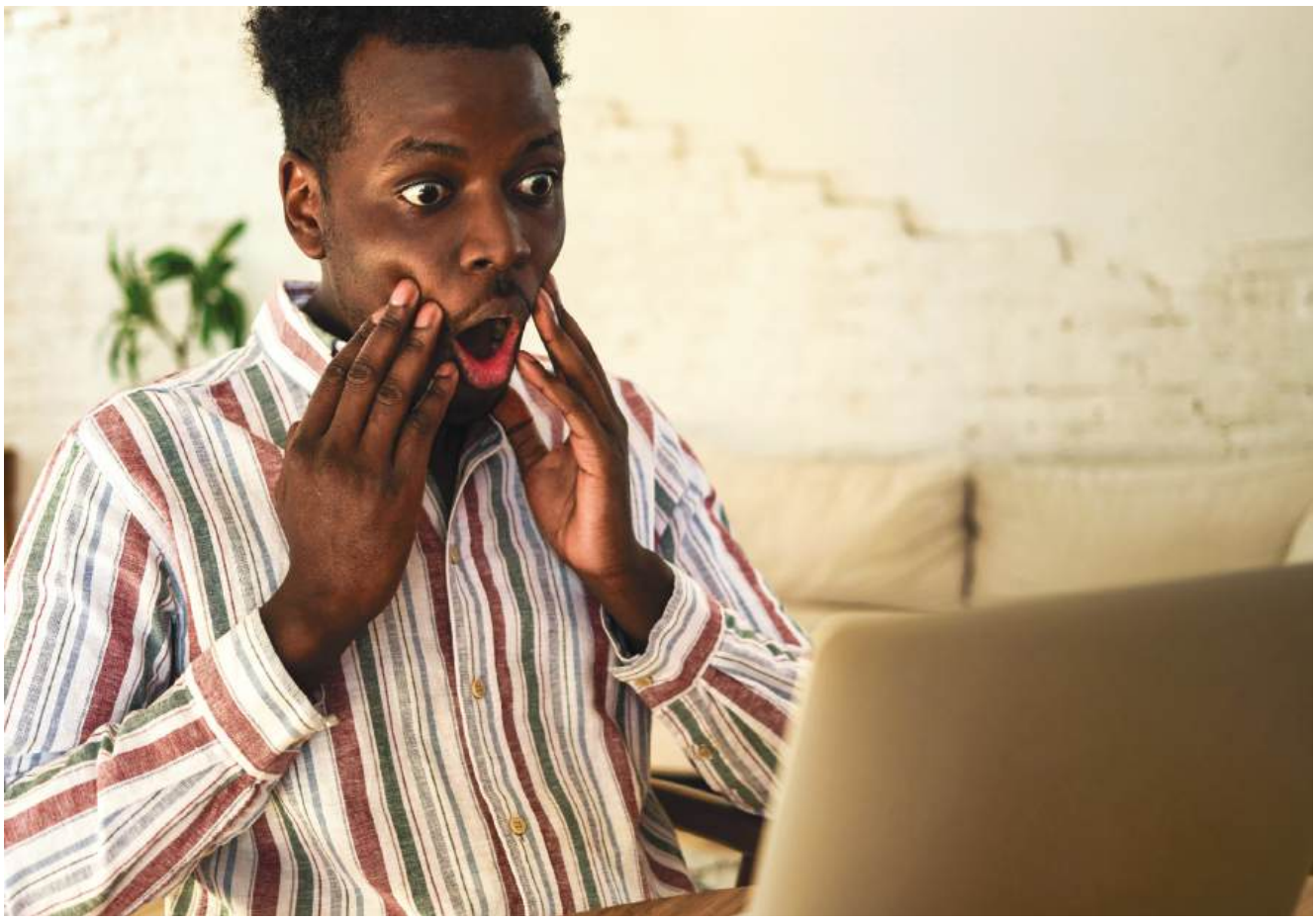
Source: Advocacy for Policy and Innovation (API) Intelligence

[73] Ibid

Similarly, the case of the Chrisland Secondary School video [74] highlights the urgent concerns regarding child pornography online and the exploitation of children in Nigeria.The widespread circulation of explicit content involving minors poses serious risks, emphasising the need for robust measures to safeguard children from online and offline exploitation. The prevalence of child pornography online allows criminal networks to exploit the internet's anonymity, increasing the risk of exposure to harmful material for young children.

The dissemination of extremist content on online platforms carries the potential to radicalise individuals, trigger violence, and deepen divisions along ethnic and religious lines, thereby challenging nation-building efforts.[75]

The impacts of these threats extend beyond physical harm. They also have the potential to negatively affect mental and emotional health, perpetuate fear and distrust between communities, and undermine national security, economic stability, and development efforts. Addressing these multifaceted challenges requires comprehensive strategies and collaboration among authorities, online platforms, and society.



[74]Tribune Editorial Board. "The Chrisland School Video." *Tribune Online*, 25 Apr. 2022, tribuneonlineng.com/the-chrisland-school-video/. Accessed 04 March 2024.

[75]Sear Richard and Johnson Neil. 'Unprecedented Reach and Rich Online Journeys Drive Hate and Extremism Globally', arXiv. November 2023. https://arxiv.org/abs/2311.08258. Accessed: 27 November 2024

2.1 Regulatory Framework for Online Harm Protection and Content Moderation in Nigeria

With the proliferation of social media and online platforms in Nigeria, the Federal Government has been prompted to make various efforts to regulate social media. However, some of these attempts have faced criticism from citizens, digital rights advocates, and civil society groups, who argue that some of these frameworks could infringe on freedom of speech and expression.[76]

In developing a new framework for protecting Nigerians from harmful content online, it is crucial to examine International law and Nigeria's current rules, regulations, and bills related to the subject matter. Understanding current jurisdiction is necessary to identify gaps and limitations to be accounted for in the proposed framework and avoid duplicity.

The following are existing laws, regulations, and bills with implications for CM and protection against harmful content online.

[76]Amnesty International (2019) 'Nigeria: Bills on Hate Speech and Social Media are Dangerous Attacks on Freedom of Expression.' Amnesty International News. December 2019. https://www.amnesty.org/en/latest/news/2019/12/nigeria-bills-on-hate-speech-and-social-media-are-dangerous-attacks-on-freedom-of-expression-2/ Accessed: 27 November 2024

2.1.1 Child Rights Act 2003

The Child Rights Act accords special protection to children as vulnerable members of the Nigerian population, outlining their rights and obligations. The Act also provides for child justice administration, care, and monitoring. Section 35 of the Act prohibits the publication and importation of harmful content (consisting of telegraphic materials portraying obscene imagery, crimes, violence, cruelty or incidents of a repulsive or horrible nature) with a tendency to corrupt children. However, the Act does not explicitly address the exploitation, posting and spreading of harmful content for and of children online, which leaves a significant gap in its coverage. This absence raises worries because the internet significantly impacts children's lives in this modern digital age.

2.1.2 Cybercrimes (Prohibition, Prevention etc) Act, 2015

The Cybercrimes (Prohibition, Prevention etc) (Amendment) Act, 2024 (first introduced in 2015) provides a comprehensive legal and regulatory framework to combat cybercrimes. It focuses on prohibiting, preventing, detecting, prosecuting, and punishing offences against critical national information infrastructure. The legislation promotes cybersecurity and safeguards various aspects of it, including computer systems, networks, electronic communications, data, computer programmes, intellectual property, and privacy rights. Some of the provisions of the Act which relate to protection against online harms include:

- The criminalisation of unlawful interception of computer systems, electronic communications, or misdirection of such communication that harms the persons concerned.
- The prohibition of access to any computer, network, or input to alter, delete, or suppress data resulting in inauthentic data intended to be used as if it were authentic or genuine.
- Prohibition of the use of computer networks to distribute or transmit child pornography or related content. The Act also prohibits solicitation of sexual engagement with a child online.
- Categorisation of cyberstalking and cybersquatting as a form of internet harassment. These may involve distributing offensive or false information about people or appropriating another person's name, business, or registered intellectual property for use on the Internet without permission. The Act considers these as threatening and potentially harmful to the life or reputation of the affected party.
- The Act also criminalises various actions related to the distribution of racist or xenophobic material through computer systems or networks. This includes threatening individuals or groups based on race, colour, descent, national or ethnic origin, or religion, as well as publicly insulting them. Again, the Act prohibits the distribution or availability of material that denies, approves, or justifies acts constituting genocide or crimes against humanity.

[49] Manila Principles on Intermediary Liability. A Global Civil Society Initiative. Version 1.0, March, 2015024.

The Cybercrimes Act is fundamentally limited by its lack of clarity on scope and procedure,[77] potentially leading to ambiguity and concerns about abuse.

Despite not explicitly using the term *"hate speech,"* the Act addresses harmful conduct online by criminalising insults based on specific characteristics. A precise definition of hate speech, avoiding extremes while adhering to best practices, is needed to improve the law's effectiveness. Additionally, although the Act's provisions relate to computer-related crimes, they do not specifically address protection from harmful content on the Internet.

2.1.3 "Social Media Bills"

In 2015, Senator Bala Na'Allah introduced a bill that aimed to *"Prohibit Frivolous Petitions and Other Matters Connected Therewith."* [78] The bill proposed penalties, including imprisonment and fines, for individuals posting abusive statements on social media or through text messages. It also required citizens to swear affidavits confirming the truth of their content before starting petitions against others. The public criticised the bill for infringing on freedom of speech, and it was eventually withdrawn by the Senate.[79]

In 2019, Senator Mohammed Sani Musa sponsored the *"Protection from Internet Falsehoods, Manipulations, and Other Related Matters Bill,"* which sought to ban prejudicial statements, grant the government authority to block internet access and prohibit hate speech, with death as a potential punishment.

This bill faced opposition, and the #SayNoToSocialMediaBill campaign gained traction on social media.[80] Citizens signed petitions and contacted their senators to protest the bill, leading to its withdrawal by the Senate.

2.1.4 Independent National Commission for the Prohibition of Hate Speeches Bill 2019 (Hate Speech Bill)

Independent National Commission for the Prohibition of Hate Speeches Bill, also known as the *"Hate Speech Bill"*[81], was one of the two bills proposed by Nigerian lawmakers to address online harmful content and fake news. The proposed bill aimed to establish a new offence, *"hate speech,"* defined as the use, publication, presentation, production, playing, provision, distribution, or direction of written or visual material threatening, abusive, or insulting. This offence is satisfied where the intent is to stir up ethnic hatred or if, given the circumstance, ethnic hatred is likely to be stirred up against any person or persons from a particular ethnic group in Nigeria. The bill prescribed severe penalties, including life imprisonment for hate speech leading to loss of life.

[77] Spaces For Change, " Amend The Cybercrimes Act Now!" https://spacesforchange.org/amend-the-cybercrimes-act-now/ Accessed June 23, 2023

[78] Taiwo-Hassan, Adebayo. "Nigerian Senate Pushes Social Media Clampdown Bill, Hits Back at Critics." Premium Times, 3 Dec. 2015, www.premiumtimesng.com/news/headlines/194386-nigerian-senate-pushes-social-media-clampdown-bill-hits-back-at-critics.html?tztc=1. Accessed 3 Mar. 2024.

[79] Editorial Board, the Guardian "The 'Anti Social Media' bill" The Guardian, 15 Dec. 2015 https://guardian.ng/opinion/the-anti-social-media-bill/ Accessed November 1, 2023

[80] Egbunike Nwachukwu. "Nigeria's social media bill will obliterate online freedom of expression." Advox. November 2019. https://advox.globalvoices.org/2019/11/29/nigerias-social-media-bill-will-obliterate-online-freedom-of-expression/

[81] Okegbile, Juliana. "Revisiting Nigeria's Legal Framework on Hate Speech and Fake News Post 2023 General Elections." Mondaq, 18 July 2023 www.mondaq.com/nigeria/social-media/1343698/revisiting-nigerias-legal-framework-on-hate-speech-and-fake-news-post-2023-general-elections.

The *Hate Speech Bill* also sought to prohibit unfair and ethnic-based discrimination, ethnic or racial contempt, hate speech and discrimination by way of victimisation by individuals or corporate bodies. Section 4 of the bill explicitly prohibits any visual or written content that may be threatening, abusive, insulting, or that involves the use of words that may trigger ethnic hatred against an individual from an ethnic group in Nigeria. The bill prescribes a punishment of life imprisonment for this offence and a penalty of death by hanging where such an act causes loss of life. While its primary objective is to promote national cohesion and integration by outlawing unfair discrimination and hate speech, the bill also received a lot of criticism. Several stakeholders have viewed it as a legislative tactic to stifle free speech.

Although the Hate Speech Bill sought to address harassment based on ethnicity, offences related to ethnic or racial contempt, and discrimination through victimisation, it is still faced with criticism, primarily for its stringent sanctions and potential impact on freedom of expression.[82] Following the first reading, concerns were raised about the extreme penalties, particularly the provision for death by hanging in cases where hate speech results in loss of life. The bill was abandoned after that.

2.1.5 Nigeria Broadcasting Code 2020 (the NBC Code)

The 6th edition of the National Broadcasting Code was published by the National Broadcasting Commission in 2016 and amended in 2020. The amendments to the NBC Code aim to curb monopolistic behaviour, promote local content, and boost advertising revenue for broadcast stations and content producers.[83] The Code sets out the minimum standards to be followed by all broadcasting operators in Nigeria, including web and online broadcasters (such as streaming platforms), who are now required to be registered with NBC.

The code prohibits any form of fake news, hate speech, inciteful comments, offensive references or general disrespect to human dignity in broadcasts. The Code also prohibits the transmission of hate speech, which is defined as "any programme, programme promotion, community service announcement or station identity, which is likely, in any circumstance, to provoke or perpetuate in a reasonable person, intense dislike, serious contempt or severe ridicule against a person or groups of people because of age, colour, gender, national or ethnic origin, disability, race, religion or political leanings."[84] Similarly, political adverts, broadcasts, and sponsored programmes that contain hate messages or speeches are prohibited. According to the 2020 code amendment, platform owners are responsible for the content on their platforms and must comply with laws and regulations related to fake news and hate speech.

Yet, this NBC code has been criticised mainly for its lack of clear safeguards against censorship and insufficient measures to ensure freedom of expression in the digital realm.[85]

[82] PWAN Analysis of The Hate Speech Bill 2019 https://www.partnersnigeria.org/pwan-analysis-of-the-hate-speech-bill-2019/

[83] Aelex Article Series. "Regulating Nigerian Content on Broadcasting Platforms: An Examination of the Amendments to the 6th Edition of the Nigeria Broadcasting Code." Aelex, June 2020. Accessed November 1, 2023 https://www.aelex.com/wp-content/uploads/2020/06/REGULATING-NIGERIAN-CONTENT-ON-BROADCASTING-PLATFORMS_-AN-EXAMINATION-OF-THE-AMENDMENTS-TO-THE-6TH-EDITION-OF-THE-NIGERIA-BROADCASTING-CODE-2.pdf

[84] Nigeria Broadcasting Commission. Nigeria Broadcasting Code (6th Edition)" NTA.ng, 2016. https://www.nta.ng/wp-content/uploads/2019/09/1494416213-NBC-Code-6TH-EDITION.pdf

[85] Jackson, Etti & Edu. "Review of Amended 6th Edition of NBC Code." jacksonettiandedu.com, August 2020. https://jee.africa/wp-content/uploads/2020/08/Review-of-Amended-6th-Edition-of-Nigeria-Broadcasting-Code.pdf

2.1.6 Digital Rights and Freedom Bill (2019)

The *Digital Rights and Freedom Bill* is intended to guard and guide today's internet users in terms of their freedom, safeguarding their rights, and protection against any form of infringement.[86] The proposed law seeks to address and reinforce the government's commitment to citizens' rights regarding internet use while emphasising freedom from unwarranted monitoring. It aims to establish a transparent framework for identifying the genuine owners of personal data and places control firmly in the hands of the individuals themselves. Notably, the legislation addresses the issue of online hate speech in Nigeria, establishing provisions to curb and combat such harmful digital behaviour.

On February 4, 2019, the leadership of the 8th National Assembly forwarded the *Digital Rights and Freedom Bill* to then-President Muhammadu Buhari for the necessary presidential assent to transform it into law.[87] The former President did not approve the bill, leading advocates to initiate efforts to reintroduce and promote it in the current 10th Assembly. The ongoing push reflects the persistent commitment of supporters to see the bill through legislative approval, underscoring its significance in addressing digital rights and freedoms in Nigeria.

2.1.7 National Information Technology Development Agency (NITDA) Code of Practice for Interactive Computer Service Platforms/Internet Intermediaries 2022

The *NITDA Code of Practice for Interactive Computer Service Platforms/Internet Intermediaries* was issued on September 26, 2022. The six-part framework aims to establish best practices, enhance the safety of Nigeria's digital ecosystem, and combat online harms such as disinformation and misinformation. The *Code of Practice* applies to all Interactive Computer Service Platforms and Internet Intermediaries operating in Nigeria. It outlines critical obligations, including swift compliance with court orders, prompt removal of unlawful content, and addressing user complaints. Additional requirements for Large Service Platforms (LSPs) include incorporation in Nigeria, human supervision of automated tools, and disclosure of advertisement reasons. Prohibitions prevent platforms from hosting illegal material, and measures against misinformation involve understanding local contexts, collaborative research, media literacy programmes, and data access for research purposes.[88]

[86] Paradigm Initiative-Reports. "Digital Rights and Freedom Bill 2019: An Analysis." Paradigm HQ, 28 July 2022. https://paradigmhq.org/report/digital-rights-and-freedom-bill-2019/

[87] Alabi, Sodiq. "Mr President, It's time to sign the Digital Rights Bill." Paradigm Initiative, 14 March 2023. https://paradigmhq.org/mr-president-its-time-to-sign-the-digital-rights-bill/.

[88] "NITDA. *National Information Technology Development Agency (NITDA) Code of Practice for Interactive Computer Service Platforms/Internet Intermediaries." nitda.gov.ng* nitda.gov.ng/wp-content/uploads/2022/10/APPROVED-NITDA-CODE-OF-PRACTIVE-FOR-INTERACTIVE-COMPUTER-SERVICE-PLATFORMS-INTERNET-INTERMEDIARIES-2022-002.pdf. Accessed 5 Mar. 2024.

While aiming to safeguard information technology systems and combat online harm, the Code faces substantial criticism. Critics believe that it potentially infringes on the right to freedom of expression by proposing interventionist principles reminiscent of the Nigeria Broadcasting Code. Its vague definition of *"unlawful content"* and lack of clarity on removal procedures continue to raise concerns about subjective interpretations and lack of judicial oversight.

The absence of broader sector consultation further exacerbates distrust, while insufficient provisions for child protection and violations of privacy policies may undermine user rights and internet standards. Moreover, the code's ambiguity regarding *"harmful"* content, morality, and state public interest leaves room for arbitrary interpretations, raising questions about who defines these terms and their statutory relevance.

2.1.8 Electoral Act 2022

The *Electoral Act of 2022* serves as a pivotal legal framework for elections in Nigeria. Section 97 explicitly prohibits all forms of sectional campaigns or broadcasts, including those based on religion and tribe, to prevent the promotion or opposition of a particular candidate.

Section 123 of the Act addresses disseminating election-related fake news, particularly regarding a candidate's withdrawal or false information intended to prejudice or promote a candidate's election chances. The section states that anyone who knowingly publishes false statements about a candidate's withdrawal or makes false statements about a candidate's character, intending to prejudice their chances or promote another candidate, and does so without reasonable grounds for belief in the statement's truth, commits an offence. The penalties include a maximum fine of N100,000, imprisonment for up to six months, or both.

However, the Act is not all-encompassing because it does not explicitly cover particular unforeseen possibilities, such as the distribution of fake election results. Furthermore, Section 125 of the Act makes it an offence for any person to act or incite others to act disorderly, with penalties upon conviction, including a maximum fine of N500,000, imprisonment for up to 12 months, or both. The term *"inciteful"* remains undefined, leaving room for any interpretation. While these provide a reasonable starting point, the effectiveness of enforcing these penalties as a deterrent to offenders appears to be limited.

2.1.9 Nigeria Data Protection Act 2023

The *Nigeria Data Protection Act* is the primary legislation on data protection in Nigeria. It provides people's rights to the safety and security of their personal information. The Act allows data subjects the right to erasure their personal information if such information is no longer necessary for the purpose it was provided or if the availability of the information in a public space (e.g. a website) adversely affects their fundamental human rights. Although Nigeria does not create a clear distinction between the right to be forgotten, which relates more to deleting personal information permanently from the internet, and the right to erasure, the latter may still be invoked as the former. The law also provides for the right to rectification, such that a data subject can request the correction of any incorrect information held about them that could be misleading.

2.1.10 Criminal Code

The *Criminal Code of Nigeria* is the cornerstone of the nation's criminal justice system. It delineates the legal boundaries within which individuals and entities must operate, addressing a comprehensive range of criminal activities, from minor to severe crimes, thereby ensuring justice and maintaining societal order. *Section 366* pertains to intimidation, criminalizing the act of coercing a person to perform an action they are not legally obligated to do or to refrain from an action they are legally entitled to do through threats of injury, harm, or damage. The effect of the provision can be extended to online intimidation, where individuals are pressured into compliance or silence through threats made via digital platforms.

Section 408 focuses on extortion, criminalizing obtaining property or any benefit through threats of harm or the exposure of secrets. This section applies to online extortion if persons are coerced into providing money or favours under the threat of releasing sensitive information or damaging reputations. Other provisions, such as criminal libel and incitement, can be applied to online harms.

2.1.11 Trafficking in Persons prohibition and Administration act

The *Trafficking in Persons (Prohibition) Law Enforcement and Administration Act of Nigeria* is a critical piece of legislation aimed at combating human trafficking. While it primarily addresses physical trafficking, its provisions are also relevant to online harms, particularly in cases where trafficking is facilitated through digital means. Section 14 of the Trafficking in Persons (Prohibition) Enforcement and Administration Act addresses the Importation and Exportation of Persons. This section criminalizes the act of importing or exporting individuals for forced prostitution or sexual exploitation. It is particularly relevant in online scenarios where digital platforms are used to facilitate such trafficking activities. Section 15 deals with the Procurement of Persons for Sexual Exploitation, criminalizing the inducement of minors or the harbouring of individuals for sexual exploitation through deception or coercion. This provision applies to online exploitation, where traffickers utilise the internet to recruit or exploit victims. Additionally, Section 21 focuses on the buying and selling of humans, criminalizing the acquisition, disposal, or possession of individuals for exploitation. These sections are also applicable to online situations where crimes that may lead to human trafficking may occur.

2.1.12 Penal Code (Northern States) Federal Provision Act

While addressing various criminal offences, including defamation and extortion, the *Penal Code* only applies to Northern Nigeria. Some key sections include Section 391, which pertains to defamation. It defines defamation as making or publishing any false statement about a person with the intent to harm their reputation. In the digital age, online defamation can also occur when individuals post false and damaging statements about others on social media, websites, or other digital platforms.

Additionally, Sections 294 and 295 highlight the crime of extortion. These sections criminalize extortion, which involves obtaining property or benefits from another person through threats or coercion. In the context of online harms, extortion can take the form of cyberbullying, sextortion, or threats to release sensitive information unless demands are met.

2.1.13 Gaps in the Regulatory Framework for Online Harm Protection and Content Moderation in Nigeria

The extant regulatory patchwork has several gaps that necessitate the introduction of robust protection from online harms regulation. These gaps are evident in the existing legal instruments and regulatory attempts, which need to be revised wholly or their approach to addressing the complexities of the digital landscape.

Controversial legislative attempts, such as the Protection from Internet Falsehood and Manipulations Bill and the Independent National Commission for the Prohibition of Hate Speeches Bill, have raised concerns over potential infringements on freedom of expression. These bills, which sought to regulate the spread of false information and hate speech, faced significant opposition due to fears that they could be used to stifle dissent and suppress legitimate speech.

To address these gaps, robust protection from online harms regulation should establish clear guidelines for online platforms regarding intermediary liability and CM and introduce measures to protect vulnerable users, especially children. An updated framework should effectively balance the need for regulation with protecting freedom of expression. Such regulation should also provide transparency and accountability mechanisms, balancing the privacy rights of individuals and risks associated with persons and enabling users to understand and challenge CM decisions. The Manila Principles for intermediary liability provides a framework for limiting intermediary liability for online content and enhancing freedom of expression.[89] The principles emphasise a shield for intermediary platforms as facilitators of conversations and enablers of innovation, a requirement for judicial authority to restrict content, and clarity in a request for restriction of content,the need for due process to restrict content and such request for restriction should comply with the test for necessity and proportionality. It is recommended that these principles be baked into the law to protect society from online harms. However, this consideration should not shield platforms where there are facts to suggest willful non-compliance to action to protect society following laid down laws or processes.

Therefore, a new and specific online harms prohibition law is essential because it will propose a structured approach to safeguarding online spaces. This is underscored by the necessity for clear regulations that preserve the fundamental rights of internet users, lean on ideas from the Manila Principles, and establish accountability for digital platforms and service providers. By aligning with this white paper's proposals, such a law will create a more secure and trustworthy digital environment.

[89] Manila Principles on Intermediary Liability. A Global Civil Society Initiative. Version 1.0, March, 2015

In conclusion, the Nigerian regulatory framework for CM requires significant enhancements to effectively manage the evolving risks associated with the digital environment. Robust protection from online harms regulation, underpinned by a commitment to human rights, user safety, and transparent governance, is essential to bridge existing gaps and foster a secure and inclusive online space for all Nigerian users.

2.2 An Opportunity to Close the Gap

This white paper examines integrating *"duty-of-care"* and co-regulatory approaches into intermediary liability models as a crucial strategy for effectively addressing the challenge of online harm within the Nigerian digital sphere. These approaches will provide the benefits of a *''duty of care''* while substantiating the role of collective and informed effort in protecting society. The approach will improve self-regulation and duty of care by including civil society and public stakeholders as critical components to ensure transparency and accountability.

2.2.1 Existing Models of Intermediary Liability

Exploring three prominent models from large democracies provides valuable insights into diverse approaches:

American Model (Section 230 of the Communications Decency Act)
Positioned as granting immunity for third-party content and moderation, this model reflects a foundational aspect of the digital landscape. The key points about the American model *(Section 230 of the Communications Decency Act)* are:

Immunity for Third-Party Content and Content Moderation
Section 230 of the Communications Decency Act protects content providers from being treated as publishers or speakers of information provided by third-party users. This reflects a vision that content providers are part of a ''dumb pipe'' system, favouring freedom of expression by protecting intermediaries and extending that protection to users.

Limited Immunity
The immunity provided by Section 230 is not unlimited. There are exceptions to federal criminal laws, illegal/harmful content, and copyright violations. These exceptions assign specific tasks to platforms, requiring them to act against particular content types.

Good Samaritan Principle
Section 230(c)(2) authorises intermediaries to moderate content and protects the removal of content done in good faith. This ''Good Samaritan'' provision exempts operators from liability when they, in good faith, remove or moderate third-party material that they deem objectionable.

Debate and Criticism

Section 230 has been the subject of much debate, with calls from Republicans and Democrats to alter or abolish the law. Critics argue the immunity granted to platforms disregards their ability to stop the spread of false information and hate speech. At the same time, proponents claim it is essential for internet freedom of expression.

In summary, the American model under Section 230 provides broad immunity for content providers, reflecting an approach favouring protecting intermediaries and user expression. However, this model has faced increasing scrutiny and calls for reform to address concerns over harmful content moderation practices.

Previous European Model (E-Commerce Directive)

Characterised by immunity alongside a "notice and takedown" approach, this model illustrates a nuanced balance between freedom of expression and content regulation. The key points about the previous European model, based on the *E-Commerce Directive,* are:

The European Union adopted the E-Commerce Directive in 2000 to establish a legal framework for electronic commerce in the EU and facilitate cross-border online transactions. The directive applies to various online services, including e-commerce, social media, and search engines. One of the key provisions of the directive is the safe harbour provision, which protects intermediaries from liability for the content they transmit or store on their platforms. Similar to the American model under Section 230, this provision is intended to encourage innovation and free expression on the Internet by limiting the legal liability of intermediaries.

However, the directive also establishes conditions under which intermediaries can be held liable for illegal content transmitted or stored on their platforms. These include cases where the intermediary has actual knowledge of unlawful activity or information on their platform or fails to remove such content once they become aware of it promptly.

Additionally, the directive provides for a "notice and takedown" procedure, which allows individuals or organisations to request the removal of illegal content from intermediaries' platforms.

In this way, the European model illustrates a nuanced balance between freedom of expression and content regulation, where intermediaries have some protection and obligations to remove content once notified.

Brazilian Model (Marco Civil da Internet):
Offering immunity for third-party content while holding content providers liable for wrongful content removal, this model exemplifies a unique approach to intermediary liability. The key points about the Brazilian model, based on the Marco Civil da Internet *(the Brazilian Civil Rights Framework for the Internet)*, are:

The Brazilian intermediary liability model, described in Article 19 of the Brazilian Civil Rights Framework for the Internet, also establishes that intermediaries are not responsible for third-party content. However, intermediaries can be required to remove content deemed illegal by court order, violate intellectual property rights, or contain unauthorised nudity.

The Brazilian model has obtained international relevance because it counts on a judicial revision to appreciate issues related to Freedom of Expression.[90] Unlike the American model, which grants content providers immunity in content moderation acts, the Brazilian model understands that these practices can violate rights and are subject to legal liability. This is why articles 19 and 21 of Marco Civil clarify the standards to be met to balance moderation of harmful content and freedom of expression, a fundamental right reinforced several times in the law.

[90] Bruna Martins dos Santos. An Assessment of the Role of Marco Civil's Intermediary Liability Regime for the Development of the Internet in Brazil. Internet Society. September 2020. https://isoc.org.br/files/Study_on_the_Marco_Civil.pdf

Chapter 3

3.0 Content Moderation and Online Harms Protection in Practice

This chapter critically examines the intricate interplay between technology, human judgement, policy interpretation, and ethical considerations of CM. Additionally, the chapter scrutinises the roles of artificial intelligence (AI) and human moderators as indispensable resources for online harm protection, emphasising the necessity for a balanced and collaborative approach. The focus extends to the pivotal role of end-to-end encryption (E2EE) in securing online communication, accompanied by a nuanced discussion of the regulatory considerations surrounding this technology. Two illustrative case studies, spotlighting the UK's Online Safety Act and the European Union's Digital Services Act, provide insights into diverse governmental approaches to CM and E2EE.

Finally, the chapter addresses Nigeria's unique context, presenting arguments for excluding E2EE from CM regulations while advocating for a holistic strategy that combines encryption preservation, collaboration with tech companies, and a mandate for platforms to contribute to online safety actively. This chapter highlights the intricate tapestry of CM and online harm protection within the contemporary digital landscape, contributing valuable insights to the broader discourse on digital governance.

"At the core of this proposed model lies a co-regulatory approach that includes civil society participation, rules obligating platforms, and transparency mechanisms for citizen involvement."

3.1 How Does Content Moderation Currently Work in Practice?

The CM landscape is dominated by three principal methodologies, each with distinct advantages and challenges.[91] The automated review model employs algorithms and is a prevalent initial defence against inappropriate content. Platforms often integrate human reviewers, drawn from their user community or through professional recruitment, to address the shortcomings of automated models. Lastly, a hybrid system combining automated algorithms and human oversight is becoming increasingly common, offering a more balanced and effective moderation strategy.

I. Manual/Human Moderation

This model relies on a platform's in-house team or Civil Society Organisations as partners to review content manually. Platforms draft a content policy that users subscribe to use the platform; moderators remove content that does not comply with this policy. This CM system is not fail-safe as there are often grey areas. Still, CM policies have been refined over the last few years, and moderators are increasingly better trained to distinguish between permissible and impermissible content.[92] Currently, this is the most accurate CM method.

[91] Gorwa, R., Binns, R. and Katzenbach, C. 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', Big Data & Society, 7(1), pp. 1-15. February 2020 https://journals.sagepub.com/doi/full/10.1177/2053951719897945

[92] Roberts S.T. "Behind the screen: Content Moderation in the Shadows of Social Media." New Haven: Yale University Press 2019. https://yalebooks.yale.edu/book/9780300261479/behind-the-screen/

The moderation process typically occurs once the content has been posted to the platform ("*post-moderation*"), although in some cases, it is vetted beforehand ("*premoderation*"). Not surprisingly, pre-moderation is less popular due to its main drawbacks it slows down the publication process, going against users' expectations for instantaneous content upload; it can result in a lack of engagement, reduced user activity, and even a loss of users; it can be resource-intensive; increases the risk of censorship; without clear content guidelines, it can lead to a lack of transparency and fairness which can erode user's trust and confidence in the platform. While pre-moderation can effectively maintain online safety, quality discussions, and brand safety, its disadvantages and inadequacy with specific needs and goals of the platform and their users lead many platforms to opt for post-moderation.

Another form of user-based moderation is where the community appoints its moderators. Reddit is a classic example of a community that uses this type of moderation, where each content channel (a "subreddit") is monitored for spam by a volunteer within that online community.[93] This can be effective for community content moderation as it requires minimal investment from the platform and leverages the benefits of a motivated subject-matter expert with an awareness of context who can respond quickly and accurately.[94]

II. Automated Moderation

As Artificial Intelligence and Machine Learning have developed, automated moderation is increasingly being deployed. It is cheaper than pure human moderation and can process large volumes of information faster than humans. Despite significant technical advancements, there are ongoing challenges with accuracy, including nuanced ethical choices, as algorithms struggle to make the "right" choice because training data from specific geographic regions and languages is scarce or not publicly available.[95]

III. The Hybrid Moderation Model

The hybrid model combines the strengths of both automated and human review systems. By integrating machine efficiency with human discernment, this approach aims to optimise the accuracy and effectiveness of CM practices. A study by Gorwa et al. explores the governance of digital platforms under this model, discussing how the blend of technology and human oversight addresses the complex challenges of online content regulation.[96]

[93] Reddit. "Moderator Code of Conduct" redditinc.com https://redditinc.com/policies/moderator-code-of-conduct

[94] Andrew, Jamie, and Ioana Burtea. "Content Moderation and Online Platforms: An Impossible Problem? Regulators and Legislators Look to New Laws." Clifford Chance, 21 June 2020, www.cliffordchance.com/insights/resources/blogs/talking-tech/en/articles/2020/07/content-moderation-and-online-platforms--an-impossible-problem--.html. Accessed 12 Mar. 2024.

[95] Gorwa, R., Binns, R. and Katzenbach, C. 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', Big Data & Society, 7(1), pp. 1-15. February 2020 https://journals.sagepub.com/doi/full/10.1177/2053951719897945

[96] Gorwa, Robert, et al. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." Big Data & Society, vol. 7, no. 1, Jan. 2020, p. 205395171989794. sagepub, journals.sagepub.com/doi/full/10.1177/2053951719897945. https://doi.org/10.1177/2053951719897945.

3.1.1 Challenges and Limitations in the Current Content Moderation Approach

The current approaches to CM face notable challenges and limitations. Most online platforms have increasingly turned to automated tools. Major platforms like YouTube, Facebook, and Twitter have expanded their AI and machine learning use. For CM purposes.

However, the accelerated adoption of these automated tools has revealed significant drawbacks. Generally, without human moderation, these tools often make errors, flagging unrelated content and hindering information sharing. Examples include Twitter's algorithm mistakenly identifying tweets containing innocuous words to require COVID-19 fact-checks and Facebook erroneously categorising posts from reputable sources as spam.

Despite technical progress, the reliance on purely automated moderation during the pandemic highlighted its inherent limitations:

I. The Fallibility of Automated Systems

Automated moderation systems, powered by AI and machine learning, promise efficiency and scalability. However, these algorithms are often hamstrung by a lack of contextual discernment, leading to the accidental censorship of benign content or, conversely, the oversight of genuinely detrimental material.[97] Machines' binary logic struggles to navigate the nuanced landscape of human communication, where context is imperative.

II. The Human Element—A Double-Edged Sword

Human moderation, though more adept at understanding context, is fraught with its own set of challenges. It is arduous, often psychologically tasking, and requires a substantial workforce to do it effectively at scale. Moreover, human judgment is inherently variable and subjective.[98] What one moderator deems offensive, another may not, leading to inconsistent application of content policies. This inconsistency can erode user trust and invite valid criticism of a platform's moderation policies.

III. Ambiguities in Content Policies

Content policies, the rulebooks guiding moderation, often contain grey areas that are open to interpretation.[99] This ambiguity can lead to inconsistent enforcement and a lack of clarity among users about what constitutes a violation. As digital platforms evolve and new forms of content emerge, these policies must be continually reassessed to draw the right lines and refined to maintain clarity and relevance.

IV. The Compromise of Post-Moderation

Post-moderation, the practice of reviewing content after publication, is favoured for its non-intrusive nature, aligning with users' real-time expectations. It is also cost-effective and efficient, thereby increasing UGC and engagement, and maintaining content authenticity. Yet, this approach displays a response to content that could potentially cause immediate damage and harm and increase inappropriate content before it is detected, assessed, and actioned against.[100]

---

[97] Gillespie, Tarleton. "The Limits of Algorithmic Content Moderation." Wired, 25 Oct. 2019. https://www.wired.com/story/the-limits-of-algorithmic-content-moderation/.

[98] Roberts S.T. "Behind the screen: Content Moderation in the Shadows of Social Media." New Haven: Yale University Press 2019. https://yalebooks.yale.edu/book/9780300261479/behind-the-screen/

[99] Suzor, Nicolas. "Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms." Social Media + Society, vol. 4, no. 3, July 2018, journals.sagepub.com/doi/10.1177/2056305118787812, https://doi.org/10.1177/2056305118787812.

[100] Lasso Moderation. 'Post-Moderation: The Pros and Cons.' Lasso Moderation. March 2023.https://www.lassomoderation.com/blog/post-moderation-pros-and-cons/ (Accessed: 27 November 2024).

V. Community Policing—The Burden of Vigilance

Platforms often rely on user reporting systems to flag inappropriate content, effectively deputising the community as content moderators.[101] This method can be effective but significantly burdens users who must police the platform. Furthermore, it can lead to biased reporting and could be more effective in smaller or less engaged communities. This method should be combined with others, including automated systems and human moderation for more prominent platforms.

VI. The Ethical Quandary of Surveillance

The increasing reliance on AI for CM raises ethical concerns about surveillance and the potential for overreach.[102] Privacy concerns are paramount, especially when monitoring private communications. The use of AI in moderation must be carefully balanced against the rights to privacy and freedom of expression.

VII. The Disparity of Resources

Effective CM requires significant resources, which may not be feasible for smaller platforms.[103] This disparity can lead to uneven enforcement across the digital ecosystem, potentially creating havens for harmful content on less-regulated sites.

VIII. The Struggle for Global Consistency

The internet's global nature demands that CM navigate different cultural norms and legal frameworks, complicating the enforcement of a consistent international standard.[104] This is especially true with platforms that provide a global service to an international user base. Platforms must balance the need for a uniform approach with respect for local nuances.

IX. The Evolutionary Pace of Online Harms

Some types of online harms evolve rapidly, necessitating continual adaptation of CM strategies to address emerging threats like deepfakes, manipulated media and sophisticated misinformation campaigns (disinformation).[105]

The current approach to CM is a complex interplay of technology, human judgment, policy interpretation, and ethical considerations. Addressing these challenges requires a multifaceted approach that combines technological innovation with nuanced human oversight, clear and evolving policy guidelines, and a commitment to ethical practices that respect user privacy and freedom of expression.

[101] Matias, J. Nathan. "The Civic Labor of Volunteer Moderators Online." Social Media + Society, vol. 5, no. 2, Apr. 2019. https://doi.org/10.1177/2056305119836778.

[102] Citron, Danielle Keats, and Frank A Pasquale. "The Scored Society: Due Process for Automated Predictions." Washington Law Review, 89(1). 8, Jan 2014. ssrn.com, 2014, papers.ssrn.com/sol3/papers.cfm?abstract_id=2376209.

[103] Gorwa, R. (2019). What is platform governance? Information, Communication & Society, 22(6), 854-871.

[104] Keller, Daphne. Internet Platforms: Observations on Speech, Danger, and Money. Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1807, 13 June 2018, www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf.

[105] Bradshaw, S., and P. Howard. "Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation." Computational Propaganda Research Project, Oxford Internet Institute, 2017, pp. 1–37. https://ora.ox.ac.uk/objects/uuid:cef7e8d9-27bf-4ea5-9fd6-855209b3e1f6

3.2 Human Moderation and AI Moderation as Tools for Online Harm Protection

The digital age has revolutionised communication, information sharing, and human interaction. However, it has also increased the prevalence of online harm, necessitating a sophisticated approach to CM. The roles of human moderation and AI moderation are pivotal, each with distinct capabilities and limitations when protecting users from harmful content.[106]

AI moderation utilises complex algorithms to oversee vast quantities of data, identifying and flagging content that may be harmful or violate platform policies. The advantage of AI lies in its ability to process information at a scale and speed unattainable by human moderators. AI systems can work continuously, applying pre-determined criteria uniformly across all content.[107] However, AI may need focused training to understand relevant context, especially regarding nuances such as sarcasm, cultural references, and idioms. This can lead to false positives, where harmless content is flagged, or false negatives, where harmful content goes undetected.[108]

Human moderation, on the other hand, excels in areas where AI falls short. Human moderators can interpret context, understand nuanced communication, and make judgments based on cultural and situational awareness. This allows for a more accurate assessment of what constitutes harmful content. However, human moderation is not without its challenges. It can be inconsistent, subject to bias, and is not scalable to the same extent as AI, making it less efficient for large-scale platforms.[109]

The most effective CM strategies employ a hybrid approach, leveraging AI's and human moderators' strengths. AI can be used for initial content filtering, handling the bulk of the workload, while human moderators can step in to make final judgments on more complex cases.[110] This collaborative approach ensures efficiency while maintaining moderation quality.

Beyond this, multistakeholder-led partnerships with civil society organisations and regulated entities are essential for enhancing online harm protection. These organisations bring expertise and can provide a valuable external perspective on CM policies and practices.[111] By working together, regulators, platforms, and civil society can develop more robust and accountable moderation systems that protect users while upholding freedom of expression.

The fight against online harm requires a nuanced approach that combines AI's scalability with human moderators' contextual understanding. By integrating these methods and fostering collaborative partnerships, online platforms can create a safer environment that respects users' rights and promotes healthy digital interactions.

The future lies in adopting a system for a mutual understanding of the landscape of online harms, establishing a 'duty-of-care' proposition, and adopting a stakeholder-led approach."

[106] Gorwa, Robert. "What Is Platform Governance?" Information, Communication & Society, vol. 22, no. 6, 11 Feb. 2019 www.tandfonline.com/doi/full/10.1080/1369118X.2019.1573914, https://doi.org/10.1080/1369118x.2019.1573914.

[107] Gillespie, Tarleton . "Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media." Yale University Press, 26 June 2018, yalebooks.yale.edu/book/9780300261431/custodians-of-the-internet/.

[108] Binns, Reuben. "Fairness in Machine Learning: Lessons from Political Philosophy." Proceedings of Machine Learning Research 81: 1-11, 2018, https://proceedings.mlr.press/v81/binns18a/binns18a.pdf

[109] Roberts, Sarah. "Behind the Screen: Content Moderation in the Shadows of Social Media." Yale University Press, 25 June 2019, yalebooks.yale.edu/book/9780300261479/behind-the-screen/.

[110] Katzenbach, Christian, and Lena Ulbricht. "Algorithmic Governance." Internet Policy Review, vol. 8, no. 4, 29 Nov. 2019, policyreview.info/concepts/algorithmic-governance.

[111] Pollicino, O. and De Gregorio, G. "Protecting Free Speech and Information in Online Platforms: A Delicate Balance." European Journal of Law and Technology, vol. 10, no. 3, 2019

3.3 End-to-end Encryption (E2EE) as a Tool for Citizen Protection

End-to-end encryption (E2EE) ensures secure communication, limiting message visibility to only the sender and receiver. It is crucial to uphold citizens' privacy during online communication, offer a protective layer for personal data, and support the fundamental principles of freedom of speech and expression.[112] Messaging platforms have increasingly adopted E2EE to secure the integrity and confidentiality of user-generated content and information.

However, there is a pressing concern about the potential misuse of this encryption technology. While it effectively shields legitimate users from unauthorised access and surveillance, it also provides a clandestine cover for wrongdoers engaged in online harm, such as cyberbullying, harassment, hate speech or other malicious activities. This duality poses a complex challenge in finding a delicate balance between preserving the privacy and security of law-abiding individuals and addressing the risks associated with malicious actors exploiting the protective umbrella of end-to-end encryption.[114]

Navigating these concerns requires a meticulous approach, considering individuals' rights to private communication and the collective responsibility to prevent and address online harm.[115] Striking this balance is a pivotal aspect of ongoing discussions surrounding the regulation and implementation of end-to-end encryption to ensure its positive impact on user privacy while mitigating potential risks associated with criminal activities conducted in secret.

3.3.1 Case Study 1: End-to-End Encryption and UK Online Safety Act 2023

The UK government's approach is intended to make the UK the safest place to be online by enacting the *Online Safety Act.* The Act has been designed to protect users' safety and privacy rights. It is deliberately tech-neutral and future-proofed to keep pace with technologies, including end-to-end encryption. It sets out a legal duty for social media companies to put in place systems and processes to tackle child sexual abuse content on their services irrespective of the technologies they use, including services using E2EE.

The Act gives OFCOM (the UK's communication regulator) the power, where necessary and proportionate, to require that a company uses accredited technology or makes best efforts to develop technology to tackle child sexual abuse on any part of its service, including public and private channels. If they fail to do so, OFCOM will be able to impose fines of up to £18 million or 10% of the company's global annual turnover, depending on which is higher.

3.3.2 Impact on the Erosion of End-to-End Encryption

Several platforms have raised concerns over the perceived erosion of end-to-end encryption (E2EE), a fundamental privacy tool in secure digital communication. These platforms argue that any attempt to weaken or compromise E2EE, whether through legislative measures or policy changes, could undermine user privacy and the security of online communications. They contend that E2EE is crucial in protecting sensitive information, ensuring that only intended recipients can access messages. Tech companies emphasise the delicate balance between privacy and security, cautioning against actions that might weaken encryption protocols and expose users to cyber threats and privacy breaches.

The *Online Safety Act* continues to raise concerns for technology companies over provisions that could undermine encrypted communications. Encrypted messaging and email services, including WhatsApp, Signal, and Element, have threatened to pull out of the UK if OFCOM requires them to install "accredited technology" to monitor encrypted communications for illegal content.[117]

Section 122 of the Act gives OFCOM powers to require technology companies to install systems that these companies and privacy advocates argue would undermine the security and privacy of encrypted services by scanning the content of every message and email to check whether they contain child sexual abuse materials (CSAM). Some intermediary platforms and activists worry that complying with the Act's provisions, particularly the potential requirement for message scanning, would compromise user privacy and introduce vulnerabilities to encrypted communications systems. The Act's power, given to OFCOM, to mandate blanket surveillance over private messaging apps is thus viewed as a significant threat to safety and privacy. Critics argue that the Act lacks safeguards for E2EE, potentially granting the government access to private communications and undermining the security measures implemented by tech companies.[118] The fear is that these measures could lead to a loss of trust in UK-based tech suppliers, harm privacy rights, and expose personal data to hackers.

European Commission (2022) Digital Services Act (DSA): 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. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065 (Accessed: 27 November 2024).

Levy, I. and Robinson, C. (2020) 'End-to-end encryption and child safety online: The UK's perspective', National Cyber Security Centre. Available at: https://www.ncsc.gov.uk/blog-post/child-safety-and-encryption (Accessed: 27 November 2024).

Article 19 (2022) The Digital Services Act and Fundamental Rights: Key Provisions and Recommendations. Available at: https://www.article19.org/resources/the-digital-services-act-and-fundamental-rights/ (Accessed: 27 November 2024).

[117] Vincent, James. "WhatsApp Says It Will Leave the UK rather than Weaken Encryption under Online Safety Bill." The Verge, 10 Mar. 2023, www.theverge.com/2023/3/10/23633601/uk-online-safety-bill-encryptionwhatsappleave. Accessed 15 Mar. 2024.

[118] "Online Safety Bill Criticism." Tuta.com. https://tuta.com/blog/online-safety-bill-criticism

3.3.3 Case Study 2: European Union's Digital Services Act (DSA) and End-to-End Encryption

The *European Union Digital Services Act (DSA)* deploys a novel approach to Intermediary liability, one that represents a comprehensive and nuanced strategy that aims to balance the protection of users activities online with the preservation of fundamental rights such as freedom of expression and privacy.

The DSA, part of the EU's digital strategy, categorises online services to tailor specific obligations to different platforms. It distinguishes between:

Very Large Online Platforms and Search Engines: These platforms reach more than 10% of the 450 million consumers in Europe and, due to their significant impact, they are subject to more stringent obligations.

Online Platforms: This category includes services that bring together sellers and consumers, such as online marketplaces, app stores, collaborative economy platforms, and social media platforms.

Hosting Services: These services, like cloud and web hosting services, store user data.

Intermediary Services include network infrastructure, internet access providers, and domain name registrars.

Within this framework, the DSA specifically addresses the issue of end-to-end encrypted (E2EE) private messaging. The Act does not classify E2EE private messaging services as online platforms because they are used for interpersonal communication between a finite number of persons determined by the sender of the communication. Instead, these services are considered 'mere conduits' as they do not host the content but merely transmit it.

The DSA is clear that providers of intermediary services should not be subjected to a general monitoring obligation concerning obligations of a general nature. The regulation emphasises that there should be no imposition of a general monitoring obligation or a general obligation for providers to take proactive measures against illegal content.[119] Moreover, the DSA introduces the concept of "due diligence obligations" for online platforms, which includes measures such as putting in place systems to detect and remove illegal content, providing users with an effective complaint mechanism, and transparency reporting on CM practices.

By categorising E2EE private messaging services as intermediary services, the DSA acknowledges the importance of encryption for the security and privacy of communications. It also recognises the technical limitations that prevent the moderation of E2EE content since it is only accessible to the sender and recipient.[120]

In summary, the EU's approach under the DSA is a model that recognises the complexity of the digital space, the diversity of services provided, and the need to protect users while respecting fundamental rights. It provides a clear and flexible framework that can be adapted to the evolving nature of online services and their challenges.

[119] Latham and Watkins. The Digital Services Act: Practical Implications for Online Services and Platforms March 2023. https://www.lw.com/admin/upload/SiteAttachments/Digital-Services-Act-Practical-Implications-for-Online-Services-and-Platforms.pdf

[120] Article 19. Recommendations for the Digital Services Act Trilogue .https://www.article19.org/wp-content/uploads/2022/02/A19-recommendations-for-the-DSA-Trilogue.pdf

3.4 Justification for an Online Harm Protection Framework in Nigeria

Prioritising online protection from harmful content is crucial in this digital age to address risks such as misinformation, disinformation and cyberbullying. Safeguarding against detrimental material protects individuals and communities and preserves mental health, societal harmony, and the integrity of online interactions and engagements. Prioritising online protection, therefore, contributes to creating a responsible and inclusive digital environment. Below are some key reasons why online harm protection should be prioritised in Nigeria.

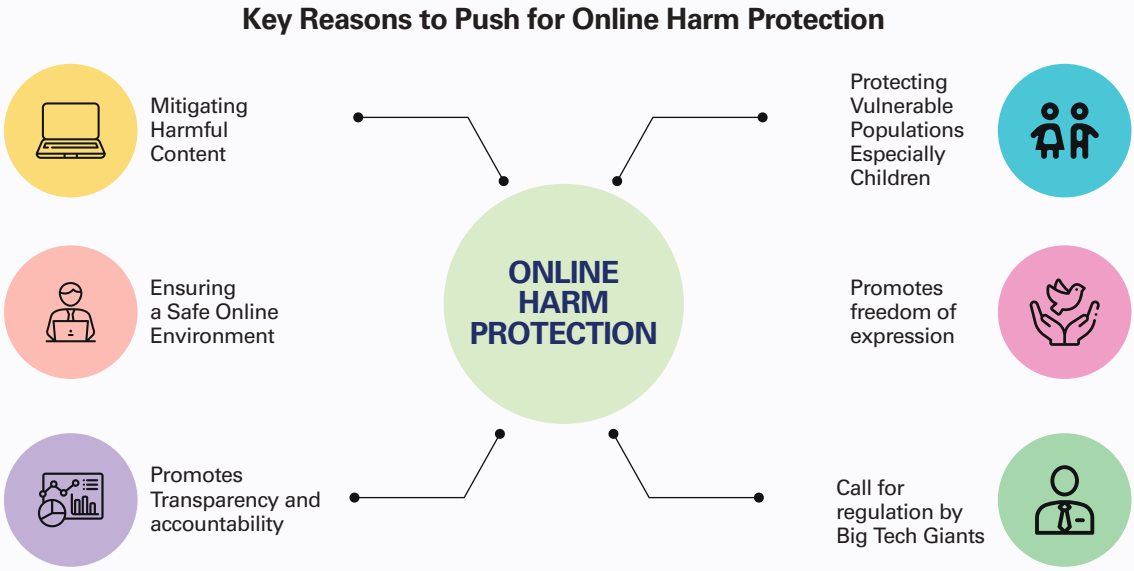


Figure 3: Key Reasons for Framework for Online Harm Protection

Source: Advocacy for Policy and Innovation (API) Intelligence

I. Mitigates Harmful Content

Regulation provides a structured framework to mitigate the spread of harmful content, ensuring a safer online environment for users and allowing various stakeholders to take responsibility for protecting Nigeria's online space.

II. Protects Vulnerable Populations, Especially Children

Regulatory measures are essential to safeguard vulnerable populations, such as minors or marginalised communities, from exploitation, harassment, and exposure to inappropriate material.

III. Ensures a Safe Online Environment

The advocacy for regulation is fundamentally rooted in the need to create a secure online space for all users. Although commendable, self-regulatory initiatives undertaken by online platforms inherently need more uniformity to grapple effectively with the evolving complexities of harmful content. Regulation, embodied in a standardised set of rules and robust enforcement mechanisms, is the cornerstone for constructing a consistent and reliable framework dedicated to user protection.

IV. Promotes Freedom of Expression

Online protection safeguards freedom of expression by creating an environment where diverse opinions can be expressed without the undue influence of harmful content or arbitrary censorship. By implementing measures that distinguish between lawful and detrimental content, online protection allows users to express themselves freely while preventing the dissemination of content that poses risks, such as hate speech, disinformation, or online harassment.

V. Promotes Transparency and Accountability

Transparency ensures that the mechanisms governing online interactions are clear and accessible. By implementing online protection measures, platforms, civil society stakeholders, and regulatory bodies can establish transparent guidelines against harmful content, reducing ambiguity and fostering a more open digital landscape. This transparency is essential to guarantee that legal content is not erroneously labelled as dangerous, addressing concerns that such mislabelling could lead to unintended content removal.

Additionally, accountability is necessary for adequate online protection. Establishing clear responsibilities for online platforms and regulatory authorities ensures that actions align with legal frameworks and ethical standards. This accountability mitigates the risk of arbitrary content removal and gives users a precise recourse mechanism in disputes. Online protective measures thus enhance the responsible conduct of platform providers and regulatory bodies.

VI. Call for Regulation by Big Tech Giants

One of the reasons why regulation is necessary is that starting in 2020, after facing probes and growing public backlash, top tech giants (Microsoft, Apple, Google, and Facebook) called publicly for new laws. Top executives of these companies are presenting global policymakers with an unusual message from an industry once antagonistic to government intervention: Regulate us.[121]

In its 2020 white paper, "*Charting A Way Forward: Online Content Regulation White Paper,*" Facebook endorsed the push for fresh regulatory frameworks governing online content. These frameworks would help platforms make decisions about online speech, striking a balance that minimises harm while upholding the essential right to free expression. The emphasis lies on safeguarding the open internet, particularly as it faces growing threats and enclosures from specific regimes.

[121] Sam Altman, the CEO of OpenAI, recently called for regulation of artificial intelligence. Kang, Cecilia. "OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing." The New York Times, 16 May 2023, www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html.

Calls for regulating Big Tech highlight the urgent need to address concerns regarding data privacy, misinformation, and monopolistic practices. However, any regulation in this sphere must adopt a balanced and collaborative approach involving citizens, governments, and Big Tech companies. While regulation is essential to protect user rights and ensure fair competition, it should not stifle innovation or undermine digital platforms' benefits. Collaborative efforts between stakeholders can foster transparency, accountability, and meaningful reforms that address the complexities of the digital landscape while preserving the dynamism of technological advancements. By engaging in constructive dialogue and considering diverse perspectives, regulations can be crafted to promote ethical practices, safeguard democratic values, and foster a healthier digital ecosystem for all stakeholders.

Finally, prioritising online protection against harmful content is vital in the digital age due to risks such as misinformation, disinformation, cyberbullying, etc.

This white paper underscores the need for a secure online environment, accountability, and protection of vulnerable populations, with a central emphasis on regulatory frameworks. The proposed regulations aim to provide standardised rules and enforcement mechanisms, addressing the limitations of self-regulation. Transparency and accountability are highlighted for clear guidelines and alignment with legal frameworks, while online protection is shown to preserve freedom of expression.

3.5 Perspectives on Excluding End-to-End Encryption from Nigeria's Protection from Online Harm Framework

As Nigeria forges ahead with its strategies to protect citizens from online harms, including (E2EE) within its scope, it has sparked considerable debate. E2EE is at the heart of private communication and is fundamental to preserving freedom of expression and privacy.

This section outlines the rationale for excluding E2EE from the proposed Online Harms Bill.

3.5.1 Upholding Freedom of Expression

Freedom of expression is a cornerstone of democracy, enshrined in the Universal Declaration of Human Rights.[122] E2EE enables individuals to communicate without fear of surveillance or censorship, fostering a climate where ideas and opinions can be [123]exchanged freely and securely. Applying CM to private messaging services would infringe upon this fundamental human right by potentially exposing private conversations to scrutiny. Importantly, Chapter IV, Constitution of the Federal Republic of Nigeria 1999 (as amended)[124] guarantees and protects the privacy of Nigerian citizens' homes, correspondence, telephone conversations and telegraphic communication.

[122] United Nations. *Universal Declaration of Human Rights,* 1945

[123] African Commission on Human and People's rights, *Declaration of Principles on Freedom of Expression and Access to Information in Africa,* 2019

[124]

3.5.2 Protecting Privacy and Security

E2EE safeguards the privacy and security of digital communications. It ensures that sensitive information, whether personal or business-related, is protected from unauthorised access. By excluding E2EE from the ambit of the proposed Online Harms Bill, Nigeria would be taking a stand to protect its citizens' privacy and uphold the security of their communications in the digital age.[125] Notably, Principle 40 of the *Declaration of Principles on Freedom of Expression and Access to Information in Africa* prohibits states from adopting laws or measures that prohibit or weaken encryption except "such measures are justifiable and compatible with international human rights laws and standards". Similarly, as Nigeria has ratified the International Covenant on Civil and Political Rights (ICCPR),[126] the country must protect the interference in citizens' privacy, family, home or correspondence.

3.5.3 Technical and Practical Limitations

From a technical standpoint, enforcing CM on E2EE services requires breaking the encryption. This would require creating vulnerabilities that could be exploited by malicious actors, thereby compromising the security of all users. It is essential to recognise these technical limitations and acknowledge that the integrity of E2EE should remain intact.[127]

3.5.4 International Precedents

Globally, there is a growing recognition of the importance of E2EE. The European Union's *Digital Services Act* (DSA) provides a framework that respects the role of E2EE in protecting user privacy. The DSA does not impose CM obligations on E2EE private messaging services, recognising them as "*mere conduits*"—a classification that should inform Nigeria's approach to its own Online Harms Protection Bill.[128]

---

[125] Kilroy, Richard. "No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State. By Glenn Greenwald, New York, NY: Metropolitan Books, 2014." Journal of Strategic Security, vol. 9, no. 3, Sept. 2016. https://doi.org/10.5038/1944-0472.9.3.1552.

[126] United Nations, International Covenant on Civil and Political rights, 1966.

[127] Abelson, Harold, et al. "Keys under Doormats: Mandating Insecurity by Requiring Government Access to All Data and Communications." Journal of Cybersecurity, vol. 1, no. 1, 1 Sept. 2015, academic.oup.com/cybersecurity/article/1/1/69/2367066, https://doi.org/10.1093/cybsec/tyv009.

[128] European Commission. Digital Services Act: EU Commission Proposes Rules for Digital Platforms, 2020. 85 Landau, Susan . "Surveillance or Security? The Risks Posed by New Wiretapping Technologies." MIT Press, 28 Jan. 2013, mitpress.mit.edu/9780262518741/surveillance-or-security/.

3.5.5 The Risk of Undermining Trust

Citizens' trust in digital platforms is contingent upon the assurance of privacy. Imposing CM on E2EE services could erode this trust, as users may no longer feel confident that their communications are private. This loss of trust could have far-reaching implications for the adoption and use of digital services in Nigeria.[129]

3.5.6 Balancing Safety with Rights

While the intent to protect citizens from online harms is commendable, balancing safety with protecting rights is imperative. E2EE plays a critical role in safeguarding these rights, and its exclusion from CM requirements would reflect a balanced approach that prioritises both security and fundamental freedoms.[130]

In the Nigerian context, we propose an approach beyond policing-specific technology by focusing on the *duty of care* for online platforms. The bill mandates platforms to actively contribute and engage in activities that promote online safety and provide accountability for internal actions and compliance with requests that affect content from civil society, citizens, or the government. Failure to fulfil these commitments should attract stringent sanctions. By requiring such consequences, the approach underscores the importance of accountability and reinforces the government's commitment to creating a safe online environment.

This combination of encryption preservation, co-regulatory practises, and the imposition of a *duty of care* reflects a holistic strategy to address the complexities of online safety in the digital landscape.

As Nigeria deliberates on a framework for the OHP bill, it is crucial to consider the negative implications of including E2EE within its scope and other state surveillance practices that can undermine freedoms and an open society. The exclusion of E2EE and other backdoor surveillance requirements from moderation requirements would align with international standards and preserve the privacy, security, and freedom of expression essential in a digital society. It would fortify the trust of Nigerian citizens in digital platforms and protect the sanctity of private communication.

[130] End-to-End Encryption. Electronic Frontier Foundation, 2021. https://www.eff.org/issues/end-endencryption

Chapter 4

4.0 A Proposed Framework for Online Harm Protection in Nigeria

The "*duty of care*" approach emerges as a strategic response to proactively address misinformation, inappropriate content, and hate speech by requiring effective monitoring and takedown responsibilities on internet content providers or online platforms. This approach emphasises proactive moderation and removing harmful content, transcending traditional notions of immunity for third-party content and positioning content providers as stewards of online safety.

Examples of regulations embracing the duty-of-care approach – including the German NetzDG, EU's *Digital Services Act,* UK's Online Safety Act, and anticipated *Brazilian Fake News Bill* – underscore the pivotal role of this method in shaping rights and obligations online.

However, as already stated, this paper also advocates for co-regulatory measures to improve transparency and accountability and mitigate the downsides of platforms' duty of care. Co-regulatory provisions will introduce the public and civil society as stakeholders to demand transparency and action, where necessary, from authorities legally.

This paper proposes a digital landscape where safety and rights coexist under a draft Online Harms Protection Bill (OHP Bill)."

4.1 Balancing Freedoms and Harms

Concerns surrounding censorship and content providers' capacity to adjudicate political speech underscore the delicate balance inherent in a duty-of-care approach. Fears of over-moderation and stifling legitimate expression highlight the nuanced challenges faced in navigating the intersection of freedom of expression and content regulation within the digital ecosystem. Therefore, legal and empowering roles for civil society and citizens will force transparency and trigger public interest litigation, where necessary, to preserve constitutional freedoms and protection.

4.2 Recommendations for Nigeria

Given the escalating concerns surrounding online misinformation and hate speech in Nigeria, adopting the duty-of-care approach emerges as a pertinent strategy. However, it is imperative to carefully calibrate duty-of-care obligations to the Nigerian context, considering the unique legal framework, technological infrastructure, and socio-cultural dynamics prevalent within the ecosystem. Tailoring duty-of-care provisions to local exigencies positions Nigeria to effectively combat online harms while safeguarding fundamental rights and fostering a thriving digital ecosystem.

The duty-of-care and co-regulatory approach would be the cornerstone for mitigating the risks of disseminating harmful online content. While combatting online harms remains paramount, the imperative lies in balancing duty-of-care obligations and preserving fundamental rights within the dynamic digital ecosystem. As policymakers chart the course forward, it is essential to ensure that duty-of-care commitments align with the overarching goal of fostering a safe, inclusive, and vibrant digital space for all stakeholders.

In this chapter, we propose that a framework for OHP, namely the "*Online Harms Protection Bill*" (OHP Bill), be drafted for enactment in Nigeria. We further proffer that the proposed law creates a co-regulatory approach to ensure transparency, responsibility, and accountability in responding to online safety issues, as a high-handed regulatory framework can significantly hamper citizens' rights, opportunities, and access.

The proposed framework comprises several vital components. Firstly, it emphasises clearly defined responsibilities for public organisations such as law enforcement and regulators, civil society, and online platforms, focusing on monitoring, responding to harmful content, and efficient complaint resolution mechanisms. Additionally, the framework mandates online platforms operating in Nigeria to establish transparent processes for addressing harmful content, with penalties for non-compliance.

To ensure effective governance and operational oversight, the framework proposes creating a Centre for Online Harms Research, Prevention, and Coordination, including representatives from public agencies, social research, and civil society. Special attention is dedicated to child online protection, with platforms obligated to prevent underage access and safeguard minors from harmful content. Together, these components aim to create a comprehensive and proactive approach to mitigating online harms in Nigeria.

4.2.1 Balanced Protection for People and Right to Privacy

Fundamentally, the proposal's crux emphasises the importance of crafting Nigeria's OHP bill with precision and consideration for the context in which content appears, distinguishing between public and private forums.

This white paper advocates for a balanced approach to OHP in Nigeria by drawing lessons from international debates emanating from legislations such as the UK's *Online Safety Act* and the EU *DSA*. This will be done under an online protection framework that protects citizens without infringing on free speech, a right enshrined in the Universal Declaration of Human Rights and the Nigerian constitution.

The framework also aims to align with international best practices and respect the technical constraints of digital communication while fostering an online environment that is safe, secure, and respectful of users' rights.

This white paper argues for excluding end-to-end encrypted (E2EE) private messaging from OHP requirements. Private conversations are akin to those in the physical world and should remain confidential, with no surveillance by the state, telecommunication providers, or messaging services. Accordingly, encryption enables privacy and human rights in the digital space.

The term "*private*," in the context of messaging, typically refers to the intended audience and the nature of the content being shared. Under this bill, a message is considered private when designed for one or a select group of recipients, with the expectation that it will not be shared beyond that audience.

Such content is classified as personal or sensitive, warranting a degree of confidentiality. The extent to which a message retains its "private" status largely depends on the platform's functionality and the user's privacy settings. For instance:

1 to 1

Direct messages between two individuals are inherently private and intended for the recipient's eyes only.

1 to a Few (Up to 5)

Messages sent to a small, closed group, such as family or close friends, generally remain private as long as all members understand and respect the confidential nature of the communication.

1 to Many

Privacy significantly diminishes once messages are sent to larger groups or public spaces. Despite the sender's initial intention, the control over who views or shares the message is reduced, and it may no longer be considered private under this bill. Ultimately, the distinction between private and public messages hinges on the sender's intent, the recipients' understanding, and the agreed-upon privacy norms within the communication channel.

Referencing best practices, the proposal suggests a framework for considering different online services when implementing OHP. This will entail categorising services and exempting interpersonal communication services, like private messaging platforms, from being considered online platforms for OHP as they are '*mere conduits*' for information.

While the government recognises that OHP is crucial for online safety, it must be implemented without compromising privacy, freedom of expression, or the integrity of E2EE.

The Nigerian Online Harms Protection Bill will incorporate provisions that:

- Clearly define moderation obligations concerning public and private online spaces, ensuring that private communications, particularly end-to-end encrypted, are exempt from content monitoring and moderation requirements.

- Uphold the principles of fundamental human rights such as free speech, freedom of association, political participation, and privacy, recognise these as fundamental human rights, and avoid overly restrictive measures that could stifle legitimate expression.

- Differentiate between online services, adopting a categorisation model to tailor moderation obligations to the service's nature and role in the digital ecosystem.

- Exclude E2EE private messaging services from moderation requirements, acknowledging their classification as 'mere conduits' and recognising the technical impossibility of moderating content inaccessible by the service provider.

- Reject proposals for technologies that undermine encryption, such as exceptional access or client-side scanning, based on expert consensus on their potential to create security and privacy risks.

4.2.2 Establishing a Regulatory Framework

The *Online Harms Protection Bill* will establish a regulatory framework for online harms protection in Nigeria. It will stipulate the roles and responsibilities of law enforcement, governmental agencies, platform operators, civil society, content developers, platform operators, and citizens. The framework will initiate a system for accountability and oversight and recognise voluntary and self-regulatory efforts. Going further, it will institute coordinated approaches and define responsibilities and accountability mechanisms to prevent individuals in Nigeria from being harmed on online platforms.

The proposed framework outlines the obligations incumbent upon stakeholders to enhance online safety for users in Nigeria and establishes clear responsibilities. It mandates a duty of care concerning illegal content and materials harmful to children while simultaneously placing obligations on platforms to safeguard users' rights to freedom of expression and privacy.

The framework will specifically regulate providers of user-to-user services, encompassing a diverse range of businesses such as social media platforms, dating apps, digital media, and online marketplaces. Generally, operators of platforms that enable user-generated content will be required to meet specified thresholds. Platforms that facilitate user content on a scale based on a gradation within the framework will bear additional responsibilities such as reporting, justification for actions taken, CM responsibilities for actions on harmful content to children, materials with significance to civic and democratic participation, and journalistic content.

Overreachingly, a robust regulatory framework aimed at combating harmful content, which includes government requests for notice and removal, should be based on four essential principles: -

i. Joint Responsibility:
Addressing illegal content represents a societal challenge where companies, governments, civil society, and users each play a part.

ii. Proportionality:
It is vital to clarify the boundaries of "control" and establish reasonable and proportionate remedial measures that intermediaries should undertake, considering the scale and nature of their services.

iii. Equity and Openness:
Require platforms or intermediaries to produce transparency reports regarding content removal and ensure users receive notifications and can contest content removal decisions.

iv. Rule of Law and Legal Clarity:
Clearly defining intermediaries' actions to meet their legal obligations, including removal duties, is essential.

Provided platforms meet the minimum legal requirements and can continue to enjoy safe harbour provisions from liability from third-party content. This will engender shared responsibilities, flexibility, and partnerships while promoting economic growth, free expression, the free flow of information, and other societal benefits.

4.2.3 Objectives of the Bill

The primary goals of the Bill are to attain key policy objectives that:

- Enhance online safety.
- Institutes proactive measures to protect children online.
- Safeguards and promotes freedom of speech in the online space.
- Strengthens law enforcement's capacity to address harmful and illegal content on the internet.
- Empower users to protect themselves better in online environments.
- Enhances society's awareness and comprehension of the landscape of harm online.
- Establishes a co-regulatory strategy involving the public and private stakeholders for transparency and accountability.

4.2.4 Scope and Applicability of the Bill

The Bill will apply to all online platforms accessible nationwide or operating within the country. The regulation will apply to tech companies, social media platforms, and online service providers.

4.2.5 Operationalising Online Protection Regulation

This section highlights the implementation mechanisms of the *Online Harms Protection Bill,* including:

Obligations for Public and Online Platforms

Clearly defined responsibilities for public organisations and online platforms, emphasising monitoring, response to harmful content, and efficient complaint resolution mechanisms.

Global Online Platforms Compliance

Mandating global online platforms operating in Nigeria to establish transparent processes for addressing harmful content, with penalties for non-compliance. This is a critical step towards ensuring a secure and responsible digital environment. Additionally, the proposed law will provide a threshold to determine the qualification and scale of human CM efforts that must be utilised on platforms, particularly during elections or other situations or happenings that may call for urgent action. The law will also stipulate transparency and reporting requirements for identifying, monitoring, and actioning harmful content in line with platform policies to help build trust between users and platforms while fostering accountability. Establishing penalties for non-compliance will serve as a deterrent, encouraging global online platforms to prioritise developing and implementing robust online harm protection mechanisms.

The *Online Harms Protection Bill* represents a pivotal step towards safeguarding online spaces from digital misconduct. Central to its provisions is the imposition of a duty of care on online platforms, requiring platforms to comply with duly enacted laws, compelling action to address egregious offences swiftly and responsibly and implementing best practices, terms and conditions to protect society from content that may be legal but harmful. Specifically, the bill will mandate platforms to fact- check and promptly remove instances of image-based sexual abuse, cyberflashing, and the creation or dissemination of deepfake pornography within a stringent but fair timeframe.

The bill aims to curb the proliferation of harmful content and protect individuals from the deleterious effects of online exploitation by imposing such measures.

This proactive approach underscores the importance of regulatory intervention in mitigating online harms and upholding the safety and well-being of Nigerian internet users.

4.2.6 Establishment of a Centre for Online Harms Research and Coordination

Considering the patchwork of laws and functions of several agencies of government (see Chapter 2) on matters relating to third-party-(digital) User Generated Content, this whitepaper proposes the creation of a coordination institution to play a crucial role in overseeing and enforcing the obligations created in the bill and coordinating the response of public agencies to protect online safety. The Centre will effectively assess and monitor adherence to the law, lead public research, and provide insights guiding further regulations or supporting healthy practices for a safe internet space. This independent oversight aims to enhance transparency, accountability, and the overall effectiveness of the regulatory framework in promoting online safety and protecting users' rights.

The proposed bill will include provisions for this Centre to operate as a research and coordination institute with active participation from law enforcement, regulatory agencies, and civil society for effective governance and operational oversight. Additionally, the centre will conduct research, publish papers, and track the evolving nature of technology's impact on third-party websites in Nigeria and the sub-region. It will provide guidance, advice, training, and insights to the government and private sector on healthy internet use.

The Centre’s governance will include memberships from relevant agencies such as the Nigerian Police, the Nigerian Human Rights Commission (NHRC), the Office of the National Security Adviser (ONSA), the National Information Technology Development Agency (NITDA), the Nigerian Communications Commission (NCC), The Federal Competition and Consumer Protection Commission FCCPC), and the National Broadcasting Commission (NBC). Civil society, academia, and social research representatives will also participate in the proposed Centre's governance. Its leadership should possess proficient research, legal, and stakeholder management skills.

A crucial challenge for the Centre will be its independence and funding. To mitigate costs and bureaucratic hurdles, leveraging existing institutions to establish the centre presents a viable solution. By tapping into established frameworks and resources, the implementation process can be streamlined while benefiting from existing expertise and infrastructure. This approach accelerates the Centre's deployment and fosters collaboration and synergy within the broader institutional ecosystem. As such, the Centre can be situated within an existing government agency with an aligned mandate and be funded through donations, gifts, or partnerships. The bill can specify the governance and operations of the centre to be independent.

The Centre will play a crucial role in overseeing and enforcing the outlined obligations in the proposed law and coordinating the response of public agencies to protect online society. The Centre would also act as a pivotal institution mediating between the imperative to combat harmful content and preserving essential freedoms in the digital realm.

4.2.7 Enhancing International Cooperation in Combatting Online Harms

In response to the escalating challenges posed by cross-border online harms such as child exploitation, terrorist content, and sextortion, there is a critical need to bolster international cooperation and establish common standards. These issues transcend geographical boundaries, necessitating coordinated efforts at a global level to combat them effectively. The proliferation of harmful content across borders underscores the imperative for enhanced cooperation and information sharing among nations. By establishing common standards and frameworks, countries can work together more effectively to address the complex challenges presented by cross-border online harms.

The paper conveys the vision for balancing rights and protections in Nigeria's digital space."

To address these challenges, the Centre will identify gaps in international cooperation and develop common standards. It will also comprehensively assess existing global frameworks, identifying areas where cooperation mechanisms for combatting online harms fall short. Building on existing best practices and experiences, the Centre will facilitate the development of common standards and guidelines for addressing crossborder issues.

Through workshops, forums, and capacity-building initiatives, the Centre will promote greater information sharing and collaboration among countries, law enforcement agencies, and relevant stakeholders, ultimately creating a safer and more secure online environment.

4.2.8 Child Online Protection Strategy

A Child Online Protection Strategy to be articulated within the Online Harms Protection Bill aims to implement comprehensive measures to safeguard minors online. A vital facet of the bill will be explicit obligations on online platforms to prevent underage access, thereby recognising the need for robust age verification mechanisms.

This strategy component will encourage platforms to use the best technology and knowhow to support age verification and identification. The Online Harms Protection Bill proposes the following provisions be included for the protection of children's safety online:

Age Assurance and Verification

All online platforms shall implement age assurance and verification mechanisms to ensure that individuals under 18 cannot access services not intended for them. Age-appropriate material should only be accessible after the user's age is verified as 18 or older, and social media sites shall put measures in place to limit access for individuals below the minimum age requirement, often set at 13 years old. In shaping our strategy, this white paper considers platforms catering to users aged 13-18. Notably, specific platforms have introduced new products tailored to this demographic, with provisions for parental guidance. These initiatives underscore a proactive approach towards ensuring the safety and well-being of young users in digital spaces.

Our strategy involves establishing collaborative partnerships with major online platforms catering to users aged 13-18 to develop robust parental supervision features. This approach advocates implementing age-appropriate content filters, time limits, and privacy settings, empowering parents to effectively manage their children's online activities. Additionally, it includes launching public awareness campaigns to educate parents about the importance of utilising these tools and fostering open communication with their children regarding online safety.

Continuous monitoring and updates of parental control features, informed by research on adolescents' digital behaviours and needs, ensure ongoing effectiveness and alignment with evolving digital trends and risks.

Through these efforts, we aim to empower parents, promote safer online experiences for young users, and foster a more responsible digital environment.

Transparency through Risk Assessments

Established larger platforms operating in Nigeria shall be obligated to publish comprehensive risk assessments regularly, outlining potential dangers and risks posed to children on their respective platforms.

Prevention and Removal of Illegal Content:

Online platforms and social media sites should actively take measures to prevent the sharing of illegal and harmful content, such as videos and images depicting child sexual abuse and exploitation. Prompt procedures should be established to eliminate such content from online platforms quickly.

Cases that would necessitate a takedown process within the bill may typically involve instances of disinformation or misinformation that may likely result in threats of violence or physical harm or the spreading of harmful content through online platforms. It is encouraged that platforms follow laid-down policies for removing content, and a judicial process should be established to review such content speedily and grant injunctions for removing this type of content. These may include:

False Information: Content that disseminates false or misleading information about significant events, public figures, or issues, leading to potential societal harm, direct violence or harm to persons, social disorder or disinformation. Also, false information may include spreading inaccurate or misleading health-related information, including false medical claims, miracle cures, or dangerous advice, potentially endangering public health and safety.

Hate Speech: Material that promotes hatred, discrimination, or violence against individuals or groups based on characteristics such as race, ethnicity, religion, gender, sexual orientation, or disability.

Cyberbullying: Harassing or threatening behaviour conducted online, including targeted attacks, intimidation, or defamation against individuals, particularly minors, leading to psychological or emotional harm.

Image-based Abuse: Sharing or distribution of non-consensual intimate images or videos (also known as "revenge porn") without the subject's consent, leading to privacy violations and emotional distress.

Manipulated Media: Manipulated media, such as deepfake videos or images, are designed to deceive viewers by presenting false or fabricated events or statements. This could cause public confusion or damage a reputation.

Misleading Advertisements: Advertisements or sponsored content that make false or deceptive claims about products or services, leading to consumer harm or fraud.

Creation of New Criminal Offences

The bill will establish new criminal offences, including but not limited to encouraging others to self-harm, engaging in trolling, purposefully targeting individuals with epilepsy by using harmful flashing content, threats to share images and sharing of deepfakes, and sending unsolicited nude photos ("*cyber flashing*").

Bereaved Parents' Right to Access Child's Data

Bereaved parents shall be granted the legal right to access their deceased child's data on online platforms, considering data protection procedural safeguards to ensure that the due process rights of individuals are respected.

Reporting Mechanisms for Parents and Children

Online platforms shall provide accessible and user-friendly reporting mechanisms for parents and children to report content that violates platform policies.

Punishment and Sanctions

The Bill shall empower regulatory bodies to impose appropriate punishments and sanctions by international human rights standards and consider global best practices for platform accountability on all online media platforms that fail to adhere to the provisions outlined in the legislation, ensuring compliance and accountability.

4.2.9 The Proposed Approach to Addressing Hate Speech

Clear Definitions
Clearly define "hate speech," "discriminatory content," and "incitement to violence," amongst other terms, to provide legal clarity and guide enforcement.

Reporting Mechanisms
Establish accessible and user-friendly reporting mechanisms for individuals to report instances of hate speech and discriminatory content to the relevant authorities or online platforms. The proposed Centre for Online Harms Research and Coordination will be pivotal in facilitating redress processes for individuals or entities affected by harmful content or online actions. Its involvement in redress processes may include:

Establishing Guidelines: Develop clear and transparent guidelines outlining the procedures for initiating and pursuing redress mechanisms. These guidelines will ensure that affected parties understand their rights, the steps involved in seeking redress, and the responsibilities of the Centre and relevant stakeholders.

Receiving Complaints: The Centre will serve as a central point of contact for individuals or entities to lodge complaints about harmful online content or actions. It will implement mechanisms to promptly and efficiently receive, assess, and document complaints.

Investigation and Evaluation: Thoroughly investigating reported cases to determine the veracity of complaints and assess the extent of harm caused. This may involve gathering evidence, interviewing relevant parties, and consulting experts to ascertain the impact of harmful content or actions.

Mediation and Resolution: The Centre facilitates mediation between affected parties, online platforms, and other stakeholders to resolve disputes and reach mutually acceptable outcomes. It may offer mediation services, promote dialogue, and provide guidance on resolving conflicts fairly and equitably.

Enforcement of Remedial Measures: Working with online platforms and regulatory authorities to enforce remedial measures, such as content removal, account suspension, or legal action, as deemed necessary to address the harm caused and prevent future occurrences.

Support and Assistance: The Centre will offer support and assistance to individuals or entities seeking redress, including access to legal advice, counselling services, or referrals to relevant support organisations. It may also guide navigating the redress process and advocating for their rights.

Timely Response Requirements

The bill will require online platforms to respond promptly to reports of hate speech and discriminatory or other harmful content, specifying a reasonable timeframe for action.

CM Provision

Online platforms must develop and implement robust CM policies prohibiting hate speech and discriminatory or other harmful content. These policies should also be regularly updated and communicated to users.

Transparency Requirements

Mandate transparency in CM practices, ensuring that online platforms provide regular reports on the prevalence of content that can cause online harm, not limited to hate speech, actions taken, and outcomes of reported cases.

Consequences for Non-Compliance

Specify consequences for online platforms that fail to adequately address reported instances of online harm, including hate speech, discriminatory content, or incitement to violence. Consequences may include fines, sanctions, or other punitive measures.

User Protection Measures

Implement measures to protect users who report hate speech, ensuring privacy and safeguarding against retaliation.

Appeal Mechanisms

Establish fair and transparent appeal mechanisms for users whose content is flagged or removed, providing an avenue for recourse in CM disputes.

4.2.10 Roles and Responsibilities of Stakeholders

Stakeholders are crucial to policy and legal development across sectors. They are central to the proposed co-regulatory approach championed by the OHP Bill. As key players, they provide vital information on the current situation, identify challenges, and propose innovative policy solutions and strategies for sector development. Their involvement is critical for informed decision-making and effective policy implementation.

Here are a few of the stakeholders who are essential in driving the passage of the OHP Bill:

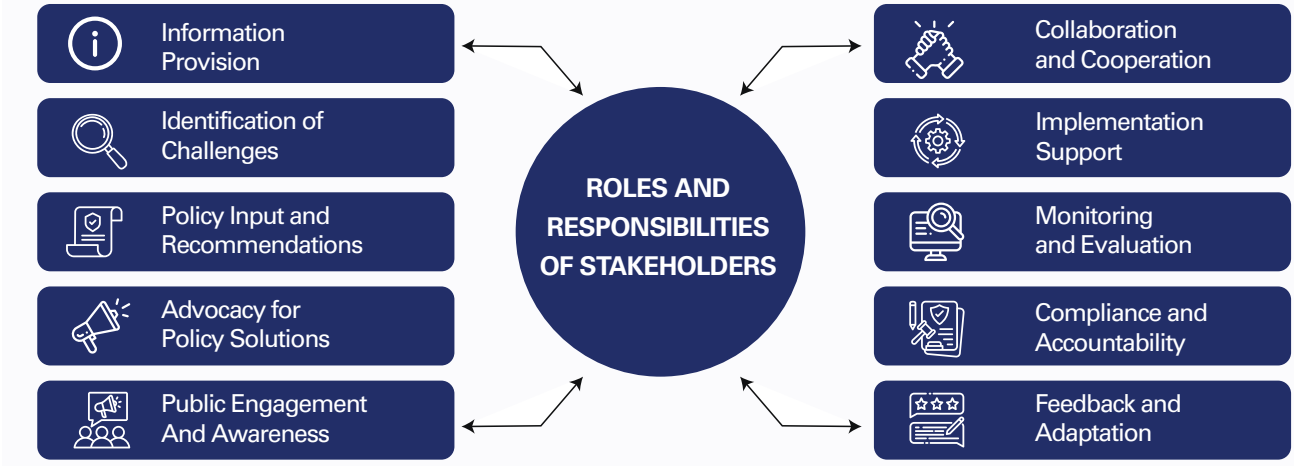


Figure 4: Roles and Responsibilities of Stakeholders in Advancing Online Harms Protection Bill in Nigeria

Source: Advocacy for Policy and Innovation (API) Intelligence

4.2.11 Duty of Cooperation and Information Sharing

Cooperation and information sharing are vital in building a secure online space. Collaboration is needed at different levels—locally, nationally, and globally. It also plays a critical role in facilitating the seamless exchange of information and is a cornerstone in the relentless pursuit of ensuring a secure online environment.

Collaboration between local entities, law enforcement agencies, and relevant organisations is vital at the state level. A cohesive approach within the national context is equally essential, involving concerted efforts from government bodies, regulatory agencies, and industry players. On the global stage, international collaboration becomes paramount as cyber threats often transcend borders. This involves sharing intelligence and best practices and coordinating responses to cyber incidents.

The establishment of a coordination Centre is a transformative initiative in this landscape. This institution will serve as a nexus for various stakeholders and provide a centralised platform for cooperation and information sharing. It will act as a multi-directional conduit, enabling seamless communication between governmental bodies, private-sector entities, academic institutions, and civil society organisations.

The Centre will serve as a hub for critical stakeholders and a dynamic hub where insights from diverse sectors converge, fostering a holistic understanding of online threats. This comprehensive approach is instrumental in developing strategies that champion online harm protection. The Centre will facilitate joint research efforts, information sharing on emerging threats, and developing proactive measures to counter online harm.

By creating a centralised repository of expertise and insights, the Centre will ensure that stakeholders are well-informed and equipped with the collective intelligence needed to stay ahead of the varying forms of online harms. This collaborative synergy will contribute significantly to the overall resilience of the digital ecosystem, making strides in fortifying the online space against a spectrum of threats.

4.2.12 Role, Responsibility and Oversight of Content Moderation Organisations

The bill will leverage the roles of fact-checkers and content moderators through the proposed legislation. The OHP Bill emphasises establishing well-equipped organisations dedicated to CM while ensuring they possess the requisite skillset for effective oversight. The Bill will further propose the standardisation of content moderators. Hence, this paper advocates for the continuous training of these professionals to keep them abreast of evolving challenges. Proper and commensurate remuneration will also be essential to attract and retain qualified individuals committed to upholding the standards of CM.

4.2.13 Role of Content Moderation Organisations:

Under the proposed bill, Content Moderation Organisations (CMOs) would be crucial guardians of truth in the digital space. Their responsibilities would include verifying the accuracy of information, debunking false claims, and providing clarity on disputed content. These organisations would collaborate with social media platforms, news outlets, and other content providers to ensure factual and unbiased information dissemination. Additionally, they would educate the public on media literacy, fostering a culture that values truth and enables individuals to discern credible sources. The law may have to set minimums for the composition of these organisations that are acceptable to all stakeholders.

CMOs would review and manage user-generated content on digital platforms to comply with legal standards and community guidelines. They would employ a holistic approach, utilising automated systems and human review processes to identify and address content that promotes hate speech, violence, terrorism, and other forms of harm.

These organisations would actively engage with stakeholders to refine content policies, ensuring transparency, equity, and the protection of freedom of expression while safeguarding users from harm.

In terms of accountability, these organisations will be expected to maintain rigorous standards of accuracy and impartiality in their operations. They must provide transparent methodologies and sources for their take-down processes and ensure that content moderation decisions are fair, consistent, and respectful of users' rights. Collaboration with authorities will involve working alongside government agencies to address emergent online threats and supporting law enforcement in investigations relating to online harms while adhering to legal constraints.

4.2.14 Comprehensive Guidelines for Protecting Digital Citizens from Granular Online Harms

To ensure comprehensive protection from online harms, a detailed set of guidelines should be developed to address specific aspects of online interactions that the overarching bill may need to cover due to their granular nature. These guidelines should supplement the proposed law, providing nuanced interpretations and practical applications for various online scenarios.

The guidelines would be structured to cover the following areas:

I. Definition of Harms:
Provide a clear and expansive list of what constitutes online harm, including less apparent forms of abuse and misconduct.

II. Scope of Application:
Clarify the extent to which private and public communications are subject to CM and the conditions under which private messages may be reviewed.

To ensure comprehensive protection from online harms, a detailed set of guidelines should be developed to address specific aspects of online interactions that the overarching bill may need to cover due to their granular nature. These guidelines should supplement the proposed law, providing nuanced interpretations and practical applications for various online scenarios.

The guidelines would be structured to cover the following areas:

I. Definition of Harms:
Provide a clear and expansive list of what constitutes online harm, including less apparent forms of abuse and misconduct.

II. Scope of Application:
Clarify the extent to which private and public communications are subject to CM and the conditions under which private messages may be reviewed.

III. User Reporting Mechanisms:
Outline user-friendly procedures for reporting harmful content, ensuring the process is accessible and efficient and respecting user privacy.

IV. Content Moderation Processes:
Detail the steps and considerations involved in content moderation, including automated tools, human review, and the balance between removing harmful content and protecting free speech.

V. Transparency and Accountability:
Require platforms to disclose moderation practices, decision-making processes, and data on handling harmful content.

VI. Appeals and Redress:
Create a standardised system for users to appeal content moderation decisions, including timelines and review processes.

VII. Protection of Vulnerable Groups:
Offer specific guidelines for protecting children, minorities, and other vulnerable populations from targeted online harms. The proposed OHP Bill will mandate guidelines that address the fast-growing threat of technology-facilitated domestic abuse. These guidelines will prioritise the safety of women and girls online, going beyond general approaches to online harms by incorporating insights from women's experiences. Furthermore, the bill will require tech companies to invest in and prioritise measures to enhance women's safety in digital spaces.

VIII. Collaboration with Law Enforcement:
Define protocols for cooperation between digital platforms and law enforcement agencies regarding illegal online activities.

IX. Education and Awareness:
Promote digital literacy programmes to help users identify, avoid, and report online harms.

X. Monitoring and Evaluation:
Implement routine assessment measures to evaluate the effectiveness of content moderation and update guidelines as necessary.

Creating and implementing these guidelines must involve collaboration among legislators, industry experts, civil society, and NITDA.

Regular reviews and updates would be essential to adapt to the evolving digital landscape and emerging forms of online harm. This proposition aims to create a dynamic and responsive appendix to the proposed bill, ensuring a safer online environment for all users.

4.3 Conclusion

This white paper recognises the dynamic nature of online interactions and the equally asymmetric scope, nature, and forms of online harms, primarily as they affect Nigerian citizens and data subjects. A hybrid approach is proposed to address this, encompassing self-regulatory efforts, civil society participation, and government oversight. This coregulatory model encourages collaboration and partnership between governments, online platforms, civil society, and citizens, fostering a shared responsibility for creating a secure and safe online environment within a duty-of-care model.

Specifically, a co-regulatory approach stipulates that platforms will initiate and execute processes to monitor harmful content defined by law proactively and will escalate illegal content through an institutional mechanism created by the proposed statute. The OHP BIll will mandate platforms to be transparent and provide proactive information on the nature, time, and actions taken regarding harmful content. Platforms will also issue periodic reports on trends and changes applied on the platform within the period in review and the impact of such changes within a stipulated time frame.

Institutionalising this regulatory framework requires the establishment of a research and coordination centre that ensures accountability through representation from multiple relevant agencies, civil society, and competent leadership with diverse skills.

Lastly, this whitepaper serves as a call to action, urging all stakeholders, including government bodies, technology companies, mainly social media and digital marketing companies, content creators, civil society organisations, and individuals, to come together in a concerted effort to shape a safer and more responsible digital landscape for Nigeria. We can only navigate the challenges posed by harmful online content and ensure the flourishing of a digital ecosystem that upholds all Nigerians' rights, safety, and security through collaborative, informed, and proactive measures.

Advocacy for Policy and Innovation (API) in Partnership with the
National Information Technology Development Agency (NITDA)