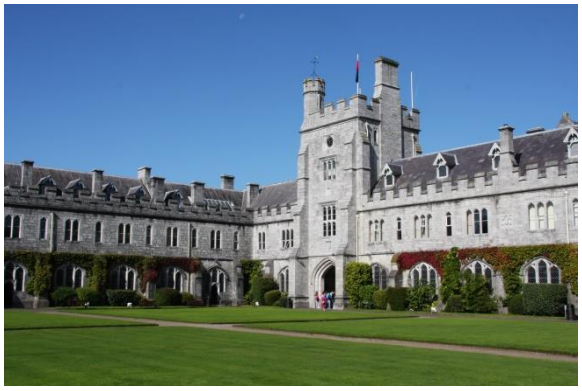
UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

School of Law
Scoil an Dlí

SHAPING A JUST SOCIETY
SINCE 1849



Legal Research Series

Working Paper No. 27

Intermediaries as co-regulators: an analysis of the EU Digital Services Act as an approach to the regulation of hate speech online

Chilombo Mukena

April 2025

CCJHR

Centre for Criminal Justice & Human Rights (CCJHR)
School of Law
University College Cork
Cork, Ireland
www.ucc.ie/en/ccjhr

CCJHR Working Papers Series: Recent Publications

1. 'Early Legal Advice and Assistance for International Protection Applicants in Ireland', Victoria Oluwatobi Isa Daniel, *CCJHR Working Paper No.16* (February 2022)
2. 'Sex Work Law in Ireland: The State's Failure to Protect Sex Workers' Human Rights under the Swedish Model', Holly O'Callaghan, *CCJHR Working Paper No.17* (February 2022)
3. 'Legislating Hate: The Growing Need for Specific Protection Against Hate Crimes and Hate Speech', Katie Fowler, *CCJHR Working Paper No.18* (February 2022)
4. 'Climate Engineering and International Law: Ignis Fatuus or Last Resort?', Johannes Mathias Schwaighofer, *CCJHR Working Paper No.19* (March 2023)
5. 'The Impact of Deforestation on Indigenous People: A Case Study of Brazil and Indonesia', Katie Place, *CCJHR Working Paper No.20* (March 2023)
6. 'An Analysis of the International, Regional and National Laws for the Substantive Rights of Internally Displaced Persons in Kenya', Bilhah Ikani Omulama, *CCJHR Working Paper No.21* (April 2023)
7. 'The Exclusionary Rule & Rights: Reforming Our Exclusionary Rule – A Comparative Analysis of the Laws of Ireland, Canada and New Zealand', Dara James Clooney, *CCJHR Working Paper No.22* (June 2023)
8. 'Child Soldiers and Juvenile Justice: Analysing the Post-Conflict Reintegration Processes for Former Child Soldiers in Sierra Leone', Estelle-Marie Casadesús Switzer, *CCJHR Working Paper No.23* (August 2023)
9. 'Leaving No One Behind: Assessing the Role of UNESCO in Promoting Adoption and Implementation of the Right to Information in Africa', Susan Juliet Agwang, *CCJHR Working Paper No.24* (July 2024)
10. 'Irregular Migrants and National Populism: The Legality of International Pushbacks by States Experiencing a Public Emergency', Anna Labadie Weeks, *CCJHR Working Paper No.25* (September 2024)
11. 'Rights, Interests, and Values in EU External Relations Law', Luigi Lonardo, *CCJHR Working Paper No.26* (October 2024)
12. 'Intermediaries as co-regulators: an analysis of the EU Digital Services Act as an approach to the regulation of hate speech online', Chilombo Mukena, *CCJHR Working Paper No.27* (April 2025)

Intermediaries as co-regulators: an analysis of the EU Digital Services Act as an approach to the regulation of hate speech online

Chilombo Mukena[1]

Abstract:

In today's digitised world, democratic discourse is largely held online with online platforms constituting the new public sphere. Intermediaries such as conduit services, caching services and hosting services such as social media platforms, web hosting platforms, content delivery networks and search engines have become synonymous with the internet itself and wield great control over the online environment. The shift of democratic discourse onto online platforms operated by these intermediaries presents unique challenges for fundamental rights and particularly the freedom of expression. This work brings into focus tensions between hate speech and the freedom of expression in the online public sphere and the role of intermediaries, which are primarily operated by privately owned businesses, in determining the acceptable limits of expression online through content moderation. The paper analyses the EUs Digital Services Act 2024 as an approach to regulating intermediary activities in that regard and whether it can provide a regulatory structure for a global approach. The paper also analyses the possibilities for an African regional approach to regulating hate speech online by applying the African moral philosophical principles of *ubuntu.* The overall aim of this paper is to highlight the need for a common approach to the regulation of content moderation activities which upholds international human rights law standards by placing clear obligations on intermediaries as duty bearers.

Key words: **freedom of expression, hate speech, intermediaries, content moderation, *ubuntu***

A. INTRODUCTION

In November 2022, the European Union adopted the EU Digital Services Act (the DSA) in an attempt to prevent illegal and harmful activities online through intermediary regulation.[2] The DSA came into force across all EU member states on 17th February 2024.[3] Among other things, the DSA establishes greater control and procedural oversight over intermediaries while guaranteeing the protection of

[1] Chilombo Mukena is a graduate of the International Human Rights Law and Public Policy LL.M at the University College Cork. She is an advocate of the Superior Courts of Zambia and her work is at the intersection of human rights law, technology and policy. This research was submitted as a dissertation for the fulfillment of a Master of Laws in August 2024, under the supervision of Dr Nessa Lynch, and has been lightly edited to reflect recent developments up to May 2025.

[2] 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 (DSA), Official Journal of the European Union I. 277/1.

[3] DSA, Article 93.

fundamental rights. The DSA recognises the need for a safe, predictable and trustworthy online environment and for persons to be able to exercise their freedom of expression and of information, and the right to non-discrimination as guaranteed by the Charter of Fundamental Rights of the European Union. It is this guarantee of rights, and particularly the freedom of expression, which this paper is concerned with. The aim of this paper is to analyse the intermediary based regulation model of the DSA as it relates to the regulation of hate speech online. This paper will further consider whether the EU model can be extrapolated and applied in the African regional context.

As of 2022, over 40% of the African population had access to broadband internet.[4] The internet has provided a platform for democratic discourse and a source of information during watershed moments on the continent. However, despite the large numbers of internet users, efforts to regulate content online have primarily been taken at state level with very few states regulating online platforms or 'over the top services'. While some states have resorted to banning specific sites altogether as a means of controlling what content citizens can share and consume,[5] others have resorted to criminalisation of hate speech in overly broad and vague terms which threaten the freedom of expression. In other states, online platforms have unbridled discretion to filter content in accordance with their own terms of service. With there being as many cultural, linguistic, historical and political contexts as there are states on the continent, this fragmented approach to regulating content online has led to inconsistent and unpredictable interpretations of what speech is allowable and uncertainty over the obligations of intermediaries.

Beginning with a brief overview of the conceptual international and regional human rights framework on the freedom of expression and its limitations, this paper discusses the role of intermediaries in combatting hate speech online, the freedom of expression concerns that arise therefrom, and approaches to regulating intermediaries. This paper further analyses the provisions of the DSA relating to hate speech by assessing the obligations placed on intermediaries and the safeguards for the protection of the freedom of expression. Additionally, this paper highlights some of the challenges of regulating hate speech online as a contemporary and nuanced problem. Finally, this work addresses how the EUs regional approach can be modelled to the African context by applying indigenous African knowledge and international human rights law to achieve a balance between curtailing hate speech and safeguarding the freedom of expression.

[4] World Bank Open Data, Individuals using the internet % of population (Africa), available https://data.worldbank.org/indicator/IT.NET.USER.ZS?most_recent_value_desc=false (date accessed: 21 May 2024).

[5] *Socio-Economic Rights and Accountability Project (SERAP) v Nigeria,* 2021 ECW/CCJ/JUD/08/21.

B. CONCEPTUAL FRAMEWORK

1. International and regional human rights law frameworks on the freedom of expression and regulation of hate speech

The freedom of expression is a central liberties tenet forming part of "a web of mutually supporting rights" such as the freedom of association, the right to vote and the freedom of opinion which are indispensable to democratic society.[6] The freedom to express oneself, to seek and share information and ideas enables democratic participation, allows informed decision-making, facilitates the search for truth and promotes democratic discourse by encouraging tolerance for diverging opinions and exposing unpopular viewpoints.[7] Despite its enabling role in democratic society, expression has the potential to threaten the enjoyment of other fundamental rights. Hate speech in particular has been shown to galvanise violence against minorities and marginalised communities and has acted as a catalyst to some of the gravest human rights violations.[8] Although there is no universally accepted definition of 'hate speech', the existing international and regional human rights frameworks provide a standard for the protection of the freedom of expression to the exclusion of speech which incites or advocates for hatred.

The Universal Declaration of Human Rights (UDHR) and the International Covenant on Civil and Political Rights (ICCPR) form the foundation of the international human rights law definitional parameters to the freedom of expression.[9] Article 19 of the UDHR provides for the freedom of opinion and expression which includes the freedom to receive and impart information through any media. Articles 19 and 20 of the ICCPR go further to establish not only the freedom of expression but its corresponding responsibilities and the permissible limitations to its enjoyment. Article 19 establishes a three-tier test by which any limitations to the freedom of expression must be provided for by law, necessary and pursue a legitimate public interest objective.[10]

---

[6] UN Human Rights Committee, General Comment No. 34, U.N. Doc. CCPR/C/GC/34 (Sept. 12, 2011) at para. 4 [hereinafter "General Comment No. 34"]; *South African National Defence Union v Minister of Defence & another,* 1999 (4) SA469 (CC) at para. 7, 8.

[7] T. I. Emerson, *The System of Freedom of Expression* (New York: Random House, 1970) at 6-7.

[8] Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, A/78/538 (18 October 2023).

[9] UN General Assembly, Universal Declaration of Human Rights 217 A (III) (1948); UN General Assembly, International Covenant on Civil and Political Rights [hereinafter "ICCPR"], United Nations, Treaty Series, vol. 999, p. 171, 16 (1966).

[10] ICCPR, Article 19 (3).

In General Comment 34, the Human Rights Committee provided guidance on the limits to the freedom of expression envisioned in Article 19 of the ICCPR.[11] With regard to legality, the Committee clarified that restrictions on speech must not be vague and must ensure, *inter alia,* that the discretion of state authorities is not without bounds.[12] Further, judicial oversight must accompany the implementation of any limitations to speech.[13] With regard to legitimacy, laws restricting the freedom of expression must pursue one of the stated aims of Article 19 (3), that is the protection of the rights or reputations of others, national security, public order, and public health or morals.[14] Finally, in order to meet the test of necessity, any limitation on the freedom of expression must be the least intrusive means and proportionate to the interest to be protected.[15] In order to determine whether a restriction is the least intrusive measure, a state must assess whether there are non-censorial methods of achieving the objectives.[16] If there are none, the state must rank available restriction options and select the least intrusive option which respects the right to speak and to share and receive information.[17] The test for proportionality requires that governments also take into account the nature of the expression and the manner in which it is disseminated.[18] These provisions apply to speech offline and online.[19]

In addition to guaranteeing the freedom of expression, the ICCPR in Article 20 places a limit to the enjoyment of this right by requiring state parties to prohibit the advocacy of hatred which constitutes incitement to discrimination, hostility or violence.[20] It is this provision that forms the basis for the prohibition of hate speech. In order for speech to be prohibited in accordance with Article 20, there must be public advocacy of national, racial or religious hatred accompanied by appropriate intent.[21] Further, the speech must create an imminent risk of hostility, discrimination or violence against persons belonging to a targeted group.[22] Restrictions imposed under Article 20 must also meet the three-part test of Article 19.[23]

[11] General Comment No. 34, *supra* note 6.
[12] *Ibid.* at para. 25.
[13] *Ibid*. at para. 40.
[14] *Ibid.* at para. 28 – 32.
[15] *Ibid.* at para. 33, 34.
[16] *Ibid*. at para. 34.
[17] *Ibid*.
[18] *Ibid.*
[19] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/32/38 (2016).
[20] ICCPR, Article 20 (2).
[21] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/67/357, para. 43 - 47 (2012).
[22] *Ibid.*
[23] General Comment No. 34, *supra* note 6 at para. 50.

The International Convention on the Elimination of All forms of Racial Discrimination (the ICERD) is also pertinent to the international legal framework on hate speech.[24] Article 4 of the ICERD requires state parties to criminalise the dissemination of ideas based on racial superiority and incitement to racial discrimination and violence.[25] The Committee on the Elimination of Racial Discrimination clarified in General Comment 35 that the acts, including speech, which are to be prohibited must possess the basic elements of intent of the speaker to influence certain conduct, incitement characterised by imminent risk that the intended conduct will result and particular harms of discrimination, violence, hatred, or contempt.[26] The Committee further emphasised that any restriction on speech should satisfy the test of Article 19 (3) of the ICCPR.[27]

On the basis of the foregoing international human rights law framework, any restriction to the freedom of expression, whether pursuant to the ICCPR or the ICERD, must meet the three-part test of legality, legitimacy and necessity. However, the Special Rapporteurs and Independent experts in the international human rights law system often find that laws regulating hate speech fail to meet at least one of these tests.[28] While some laws are vague and run the risk of being used to silence critical opinions, others are inspired by governmental objectives that fall short of the legitimacy test.[29]

While the international human rights law system has arguably established principles for the scope of the freedom of expression, there is some variance at regional level in the content and scope of the right. Furthermore, regarding hate speech, there are noteworthy differences in the tests applied to determine the lawfulness of limitations. For purposes of this paper, the European and African regional human rights systems will be considered.

(a) The European Human Rights System

The European Convention for the Protection of Human Rights (the ECHR) and the Charter of Fundamental Rights for the European Union (the EU Charter) guarantee the freedom of expression which encompasses the freedom to hold opinions, receive and impart information and ideas without

[24] UN General Assembly, *International Convention on the Elimination of All Forms of Racial Discrimination*, U.N.T.S Vol. 660, 21 December 1965 [hereinafter "ICERD"].
[25] ICERD, Article 4.
[26] Committee on the Elimination of Racial Discrimination, General Recommendation No. 35, Combatting Racist Hate Speech, U.N. Doc. CERD/C/GC/35 (Sept. 26, 2013).
[27] *Ibid.* at para. 12.
[28] E. Aswad and D. Kaye, "Convergence & Conflict: Reflections on Global and Regional Human Rights Standards on Hate Speech" (2022) 20 *North Western Journal of Human Rights* 165 at 183.
[29] *Ibid.* at 183.

interference by public authority.[30] Although the rights guaranteed by both instruments are largely overlapping, the EU Charter applies only to member states of the European Union (the EU) and is interpreted by the Court of Justice of the European Union while the ECHR applies to all member states of the Council of Europe including those outside the European Union and is interpreted by European Court of Human Rights.[31]

Unlike the ICCPR and the ICERD, the ECHR and the EU Charter do not contain an outright ban on hate speech. Under the ECHR, any restrictions to the freedom of expression must be legal, necessary in a democratic society and pursue the aims of national security, territorial integrity or public safety, maintenance of order, protection of health or morals, protection of the reputations or rights of others, preventing disclosure of confidential information, and maintaining judicial authority and impartiality.[32] The Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of racist and xenophobic nature committed through computer systems is also of particular importance to freedom of expression online and the prohibition of speech which incites hatred.[33]

Through its jurisprudence, the European Court of Human Rights (the ECtHR) has established two categories of hate speech: the 'gravest' forms of hate speech which are excluded from protection under the ECHR pursuant to Article 17 and 'less grave' forms of hate speech which do not fall entirely outside the protection of the ECHR but which states may limit.[34] The former category of speech comprises speech which is aimed at the destruction of any of the rights set out in the ECHR. The latter category of speech may be determined by states within a 'margin of appreciation', provided the tests of legality, legitimacy and necessity are met.[35]

The approach of the ECtHR in determining the lawful limitations to the freedom of expression departs from the international human rights law standard of the United Nations (UN) system described above on several key points. Firstly, by allowing states a 'margin of appreciation' to determine what qualifies

[30] Convention for the Protection of Human Rights and Freedoms, Council of Europe Treaty Series, CETS No. 5, 4.1.1950) [hereinafter "ECHR"] Article 10; Charter of Fundamental Rights for the European Union (Official Journal of the European Union 2012/C 326/02) Article 11.

[31] Charter of Fundamental Rights for the European Union (Official Journal of the European Union 2012/C 326/02) Recitals; Protocol No. 11 to the Convention for the Protection of Human Rights and Fundamental Freedoms, restructuring the control machinery established thereby (ETS No. 155).

[32] ECHR, Article 10 (2).

[33] Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of racist and xenophobic nature committed through computer systems, ETS 189 – Cybercrime (Additional Protocol), 28.I.2003, Article 2.

[34] *Lilliendahl v Iceland,* no. 29297/18, para. 33-40.

[35] *Tagiyev v Azerbaijan*, App. No. 13274/08 (December 5, 2019).

as illegal hate speech, the ECtHR departs from the position of the Human Rights Committee which rejected the application of a margin of appreciation in relation to the freedom of expression.[36] Secondly, the ECtHR has upheld laws that would otherwise be considered vague under the UN system, opting to consider vagueness as part of the test for proportionality rather than legality thereby watering down the three-part test.[37] The ECtHR has also upheld the defence of religious sensibilities as a legitimate ground for restricting speech, a position which has been rejected by the Human Rights Committee.[38] Finally, the ECtHR does not require states to demonstrate that their restrictions are the least intrusive measure,[39] or that there is intent to cause harm or that harm resulting from hate speech is imminent.[40] 'Wanton denigration' and insulting speech or ridicule may therefore be limited by states.[41]

(b) The African Human Rights System

Article 9 of the African Charter on Human and Peoples' Rights (the African Charter) guarantees the freedom to receive information and freedom to express and disseminate opinions within the law.[42] As with the ECHR, the African Charter does not set out an outright ban on hate speech within its text. The primary monitoring and enforcement mechanism of the African Charter is the African Commission on Human and Peoples' Rights (the ACHPR).[43] The African Court on Human and Peoples' Rights (the ACtHPR) complements the protection mandate of the African Commission.[44] In addition to the protection mechanisms under the African Charter, the African Union recognizes 8 sub-regional economic communities established under treaties independent of the African Charter.[45] Of the 8 communities, the Community Court of Justice of the Economic Community of West African States (the ECOWAS Court) and the East African Court of Justice (the EACJ) are the most noteworthy in their contributions to the regional human rights jurisprudence. The instruments establishing both Courts recognise the need to protect human rights in accordance with the African Charter.[46] Although the

[36] General Comment No. 34, *supra* note 6 para. 36.

[37] Aswad and Kaye, *supra* note 28 at 195; *Terentyev* Application No. 10692/09 para. 30, 42.

[38] *Otto-Preminger-Institut v Austria*, 295 Eur. Ct. H. R. (1994); General Comment No. 34, *supra* note 6 at para. 48.

[39] ECtHR Fact Sheet on Hate Speech 2020.

[40] *Soulas and others* App. No. 15948/03.

[41] *Savva Terentyev*, App. No. 10692/09 para. 68.

[42] African Charter on Human and Peoples' Rights (Banjul Charter) (1982) 21 I.L.M. 58, Article 9.

[43] *Ibid*, Article 30.

[44] Protocol to the African Charter on Human and People's Rights on the Establishment of an African Court on Human and Peoples' Rights (1998) Article 3, 4.

[45] S. T. Ebobrah, "Application of the African Charter by African Sub-Regional Organisations: Gains, Pains and the Future" (2012) 16 *Law, Democracy and Development Journal* 49 at 52.

[46] Treaty for the Establishment of the East African Community (2000) Article 6; Revised Treaty of the Economic Community of West African States (1993) Article 4 (g); *Korua v Niger*, ECW/CCJ/JUD/06/08.

jurisprudence on freedom of expression in the context of hate speech within the African regional human rights system is limited, both the regional and sub-regional bodies have interpreted the guarantee of Article 9 of the African Charter with varying results.

The ACHPR has interpreted the guarantees of the African Charter as requiring that any restrictions to speech be legal, legitimate and necessary within the terms of Article 19 of the ICCPR.[47] With regard to legality and legitimacy, the ACHPR has further determined that vague laws do not meet the test for legality and that the grounds for legitimate restrictions to speech are limited to the protection of rights, collective security, morality and common interest.[48] The EACJ also applies the three-part test to determine whether a restriction on speech can be upheld and adopted the position of the Human Rights Committee regarding the application of the least intrusive mode of restriction.[49]

In 2019, the African Commission adopted the Declaration of Principles on Freedom of Expression and Access to Information in Africa endorsing UN interpretations of the freedom of expression.[50] The Declaration includes a tripartite test mirroring Article 19 of the ICCPR and a 'least intrusive measure' test which are not in the text of the African Charter. In contrast to the approach of the ECtHR, the Declaration does not make mention of a 'margin of appreciation'. In applying the Declaration, the ECOWAS Court determined that its provisions posit the parameters of the exercise of the freedom of expression under the African Charter and a failure to comply with the principles therein contravenes Article 9 of the African Charter.[51]

2. **Freedom of expression in a digital age**

Technological advancement over the last few decades has introduced new dynamics to the freedom of expression. The international instruments discussed above were adopted in a time where expression was largely geographically limited, attributable to identifiable speakers and subject to the

[47] *Kenneth Good v the Republic of Botswana,* Communication 313/05 (2010), para. 187.

[48] *Social and Economic Rights Action Centre & Another v Nigeria*, Communication 15/96 (2001), para. 165.

[49] *Media Council of Tanzania and 2 others v Attorney General of the United Republic of Tanzania*, Ref. No. 2 of 2017, Para. 66-90.

[50] African Commission on Human and Peoples' Rights, *Declaration of Principles on Freedom of Expression and Access to Information in Africa,* adopted at the Commission's 65th Ordinary Session in November 2019, Banjul, The Gambia (2019).

[51] *Incorporated Trustees of Expression Now Human Rights Initiative v Federal Republic of Nigeria*, ECW/CCJ/JUD/37/23, Para. 51.

regulations of individual states. With the benefit of the internet, expression online is instant, multijurisdictional, and enjoys indefinite discoverability and anonymity.[52]

In Barlow's famous 1996 Declaration of the Independence of the Cyberspace, he describes the internet as a realm outside the authority of state power and beyond the reach of traditional legislation where 'legal concepts of property, expression, identity, movement and context' did not apply.[53] Nearly two decades later, the cyberspace is a networked public sphere regulated by laws and contracts where states, once intent on exercising their sovereignty, have resigned to the fact that they cannot exercise complete control.[54] In comparison to traditional forms of media such as newspapers, expression online is seemingly beyond the reach of states as control over these spaces is exercised by the intermediaries which host such expression.[55]

The internet and online media has experienced a shift from concerns over state power to concern over the "commercialisation, commodification, and propertisation" of cyber space.[56] Private companies driven by profit have wide discretion and responsibility over the acceptable limits of speech online.[57] Through architecture, software and terms of use, intermediaries are primarily responsible for curating the online environment by filtering the speech of users.[58] In a process referred to as content moderation, intermediaries screen user-generated content to determine its compliance with legal standards and user terms.[59] Content found wanting may be removed and user accounts suspended or terminated. Through content moderation processes, intermediaries make legal decisions about speech online considering their own commercial interests with limited, if any, judicial oversight.[60]

---

[52] *Delfi AS v. Estonia,* ECtHR Grand Chamber Application no. 64569/09, at para. 148.

[53] John Perry Barlow, *A Declaration of the Independence of Cyberspace* (1996) available https://www.eff.org/cyberspace-independence

[54] A. Murray, *Information Technology Law: The Law and Society,* 4th ed. (Oxford: Oxford University Press, 2019) at 34.

[55] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/38/35 (6 April 2018) at 3.

[56] F. R. Jørgensen, "Framing Human Rights: Exploring Storytelling within Internet Companies" (2017) 21 *Information, Communication & Society* 340 at 343.

[57] T. Gillespie, "Platforms are Not Intermediaries" (2018) 2 *Georgetown Law Technology Review* 198 at 200.

[58] F. R. Jørgensen, *Framing the net: The Internet and human rights* (Cheltenham: Edward Elgar, 2013).

[59] S. T. Roberts, "Content Moderation" in L.A Schintler, C.L McNeely (eds.) *Encyclopedia of Big Data* (Springer, 2022).

[60] C. C. V. Machado and T. H. Aguiar, "Emerging Regulations on Content Moderation and Misinformation Policies of Online Media Platforms: Accommodating the Duty of Care into Intermediary Liability Models" (2023) 8 *Business and Human Rights Journal*, 248.

The regulation of hate speech by intermediaries through content moderation raises three key concerns for the protection of the freedom of expression. Firstly, through their terms and conditions and community guidelines these private entities have a direct impact on what speech is permissible.[61] Online platforms have adopted different definitions of what constitutes 'hate speech' which are enforced to varying degrees by different platforms.[62] For instance, while Meta takes a tiered approach to 'hateful conduct' which it defines as including expressions of contempt, cursing and calls for exclusion on the basis of its categories of protected characteristics,[63] X (formerly Twitter) prohibits 'direct attacks' including through hateful references, incitement, slurs and tropes, dehumanisation and hateful profiles.[64] These disparities have potential to stifle legitimate speech such as legitimate debate and diverse views which, while not outside the ambits of the protection of international human rights law, run afoul of a particular intermediary's terms and conditions. These terms and conditions could also be relied on by states to clamp down on critical views and suppress dissenting opinions.[65] What is determined to be hate speech and therefore regulated can have dire impacts on democratic discourse and fundamental rights.

Secondly, the processes by which impermissible speech is identified and removed from the online public sphere have largely been unknown and based on unclear terms.[66] Intermediaries do not set out how their terms and conditions, community guidelines and filters translate into processes of identifying, removing and blocking content which raises concern over the possibility of these processes to be used for censorship with no recourse for users.[67] Without procedural transparency, consideration of user rights or due process, content moderation has potential to become "a new privatized digital form of prior restraint over public speech."[68]

Finally, through algorithmic design intermediaries control what content is visible online.[69] While the visibility of any particular content online may be influenced by user choices and legal limitations such

[61] B. Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-based Approach to Content Moderation" (2020) 43 *Fordham International Law Journal* 939 at 944.
[62] *Ibid* at 960.
[63] Hateful Conduct. In Facebook Community Standards, Transparency Centre, Meta. Available https://transparency.meta.com/en-gb/policies/community-standards/hate-speech (date accessed: 20 April 2025).
[64] Hateful Conduct, available X's policy on hateful conduct | X Help (date accessed: 20 April 2025).
[65] Sander, *supra* note 61 at 960.
[66] J. Venturini et al, *Terms of Service and Human Rights: An Analysis of Online Platform Contracts* (Editora Rivena, 2016) at 58.
[67] R. F. Jørgensen and L. Zuleta, "Private Governance of Freedom of Expression on Social Media Platforms: EU content regulation through the lens of human rights standards" (2020) 41 *Nordicom Review* 51 at 59.
[68] Sander, *supra* note 61 at 944.
[69] Gillespie, *supra* note 57 at 198.

as liability laws, as profit driven entities, intermediaries have corporate imperatives to "promote engagement, increase ad revenue, and facilitate data collection."[70] Some intermediaries such as social media platforms operate on business models which prioritise advertising and in some cases incentivise and promote extreme and inflammatory content such as hate speech to keep users engaged and drive up sales.[71]

Online speech and its regulation have far reaching real-life consequences for democracy and human rights. When intermediaries fail to act against speech which incites hatred, it has the potential to act as an accelerant for violence by lending legitimacy to such speech.[72] Social media platforms have been used to spread hate speech and disinformation in particularly fragile democracies during elections, swaying public opinion and igniting violence.[73] The online platform Facebook was used by military and political leaders to spread hate and incite genocidal violence against the Rohingya in Myanmar.[74] The United Nation's fact-finding mission on Myanmar reported of the online platform Facebook that it was "a useful instrument for those seeking to spread hate in a context where for most users Facebook *is* the Internet" despite its terms of service which clearly prohibit hate speech.[75]

With billions of users worldwide, the sheer size and scale of intermediaries raises concern for the unique impact they have on the freedom of expression.[76] In 2015, the UN Special Rapporteur on Freedom of Expression rightly determined that private actors and their role in shaping the use of the internet is one of the most pressing issues of the digital age.[77] In 2018, a Committee of experts on internet intermediaries conducted a study on the human rights dimensions of automated data

[70] *Ibid* at 199.

[71] A. Keen, "The 'Attention Economy' Created by Silicon Valley is Bankrupting Us", *Techcrunch* (28 January, 2020); D. Ghosh, "Facebook's Oversight Board is Not Enough" *Harvard Business Review* (16 October, 2019); Ronald Deibert, "Three Painful Truths About Social Media" (2019) 30 *Journal of Democracy* 25, 31 – 34.

[72] D. K. Citron, "Civil Rights in Our Information Age" in S. Levmore and M. C. Nussbaum (eds.), *The Offensive Internet: Speech, Privacy and Reputation,* (Cambridge: Harvard University Press, 2010) at 37.

[73] Platforms' Election Interventions in the Global Majority are Ineffective, *Mozilla* (27 February 2024) https://foundation.mozilla.org/en/blog/mozilla-research-platforms-election-interventions-in-the-global-majority-are-ineffective/

[74] S. Frenkel and C. Kang, *An Ugly Truth: Inside Facebook's Battle for Domination* (Harper, 2021).

[75] Report of the Independent International Fact-Finding Mission on Myanmar, U.N. Doc A/HRC/39/64 (12 September 2018) at 14.

[76] Jørgensen and Zuleta, *supra* note 67 at 56.

[77] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/32/38 (2016).

processing and arrived at a similar conclusion expressing concern over the automated processes used by intermediaries to filter and remove content online without a clear legal basis.[78]

While states remain the primary duty bearers in international and regional human rights law, they no longer have sole control over the means by which their citizens may express themselves.[79] The obligations of states regarding the freedom of expression in international and regional human rights instruments discussed above do not extend to the private companies which operate the majority of intermediaries.[80] In efforts to apply human rights standards to business activities and guide corporate responsibility, the Guiding Principles on Business and Human Rights (the UNGP) were developed by the Special Representative of the Secretary General of the United Nations and endorsed by the Human Rights Council as the "authoritative global reference point for business and human rights".[81] Although non-binding, the UNGP establishes best practices for all businesses, chiefly the responsibility to respect international human rights by avoiding infringement of human rights, avoiding causing adverse impacts to human rights through their own activities and mitigating adverse human rights impacts directly linked to their operations regardless of their size and operational context.[82] These principles apply to privately owned intermediaries and their activities.

3. **Conclusion**

The international and regional framework provides a clear standard for the protection of the freedom of expression and the acceptable limitations to such freedom notwithstanding the lack of a universally accepted definition of 'hate speech'. However, contemporary issues arising from the exercise of the freedom of expression online, and particularly the role of intermediaries such as online platforms, warrant further consideration of how these human rights standards are upheld in law and practice. The gatekeeping functions of intermediaries and the regulation of these intermediaries is a crucial part of human rights today as it affects the way in which fundamental rights and freedoms are enjoyed

[78] Council of Europe, Committee of experts on internet intermediaries, Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications, at 19.
[79] J. Viljaen, "Combating hate speech online", in M. Susi (ed.) *Human Rights, Digital Society and the Law: A Research Companion* (Oxford: Taylor & Francis Group, 2019) at 233.
[80] M. K. Land, "Regulating Private Harms Online: Content Regulation under Human Rights Law" in F. R. Jørgensen (ed.), *Human Rights in the Age of Platforms* (MIT Press, 2019) at 287.
[81] UN Human Rights Council, Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises on Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy: A Framework for Business and Human Rights", U.N. Doc. A/HRC/8/5 (2008) [hereinafter "UNGP"].
[82] *Ibid.* principles 11, 12, 13 and 14.

online and how democratic values are upheld.[83] It is concluded that the imposition of clear human rights obligations on intermediaries, which are largely privately owned and operated, is imperative to the effective exercise of rights.

C. REGULATION OF ONLINE HATE SPEECH IN THE EU DIGITAL SERVICES ACT

1. Historical context and development of the EU response to regulating hate speech online

The EUs regional response to hate speech began at the first Summit of Heads of State and Government of the member states of the Council of Europe in 1993 where the European Commission against Racism and Intolerance (the ECRI) was established to address concerns over the growing phenomena of racism, intolerance and discrimination.[84] The establishment of the ECRI was followed by the 1996 Joint Action to combat racism and xenophobia by which states agreed to undertake measures to ensure that behaviour such as public incitement to discrimination, violence or racial hatred are punishable as criminal offences.[85]

In 2008, the EU Council adopted the Framework Decision on combating certain forms of expression of racism and xenophobia by means of criminal law (the Framework).[86] The Framework provides an approximation of laws and regulations for certain serious manifestations of racism and xenophobia which ought to be criminalised in all EU member states and met with "effective, proportionate and dissuasive penalties."[87] "Hatred" in the Framework refers to "hatred based on race, colour, religion, descent or national or ethnic heritage."[88]

Article 1 of the Framework set out intentional conduct which ought to be criminalised and included publicly inciting violence or hatred against a group of persons or a member of such a group defined by reference to race, colour, religion, descent, national or ethnic origin through public dissemination of material. It also includes publicly condoning, denying or grossly trivialising crimes of genocide,

[83] Machado and Aguiar, *supra* note 60 at 251.

[84] Council of Europe, Vienna Declaration, Appendix III Declaration and Plan of Action on combating racism, xenophobia, antisemitism and intolerance (09/10/93).

[85] Joint Action of 15 July 1996 adopted by the Council on the basis of Article K.3 of the Treaty on European Union, concerning action to combat racism and xenophobia 96/443/JHA.

[86] Council Framework Decision 2008/913/JHA of 28 November 2008 on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law, 2008 O.J (L 328) 55.

[87] *Ibid,* at para. 5.

[88] *Ibid,* recital 9.

crimes against humanity and war crimes as defined by the Statute of the International Criminal Court where such conduct is carried out in a manner likely to incite violence or hatred.[89]

Despite these efforts, in 2016 the ECRI reported a sharp increase in xenophobic hate speech and nationalist rhetoric due, in part, to the rise in the number of migrants entering the EU.[90] That same year the Commission, fearing that anti-immigrant rhetoric would lead to terror attacks, adopted a voluntary Code of Conduct in conjunction with the privately owned online platforms Facebook, Twitter, YouTube and Microsoft to counter illegal hate speech online.[91] The Code of Conduct has since been joined by Tiktok, Instagram, Snapchat, Dailymotion, Jeuxvideo.com, LinkedIn, Rakuten Viber and Twitch.[92] By 2019, the Code of Conduct covered 96% of the EU market share of online platforms which are susceptible to hateful content.[93]

The Code of Conduct adopts the definition of hate speech provided in the 2008 Framework and seeks to complement legislation against hate speech by ensuring that hate speech online is acted upon by online intermediaries and social media platforms. Under the Code of Conduct, intermediaries undertake public commitments to establish guidelines for users clarifying the prohibition of hate speech and establish notice-and-action procedures to detect and remove illegal hate speech. The signatory companies committed to reviewing the bulk of valid notifications within 24 hours of receipt and to action such notifications by removal of content where necessary. As the Code of Conduct is voluntary, there are no enforcement mechanisms. The Commission and the online platforms and intermediaries agreed to periodically assess the commitments in the Code of Conduct and the progress on curtailing hate speech online.

The proposal for legislation to consolidate the various frameworks and establish a common standard for the regulation of various forms of illegal speech was first made in 2020. After negotiations between the EU Parliament, Council and Commission, the consolidated text of the DSA was voted on in July 2022 and came into force across all EU member states in February 2024.

[89] *Ibid,* Article 1.

[90] ECRI General Policy Recommendation No. 16: Safeguarding irregularly present migrants from discriminations (10 May 2016).

[91] Code of conduct on countering illegal hate speech accessed (30 June 2016).

[92] E. Aswad, "The Role of U.S. Technology Companies as Enforcers of Europe's New Internet Hate Speech Ban" (2016) 1 *Columbia Human Rights Law Review Online* 1.

[93] Information Note on Progress on combating hate speech online through the EU Code of Conduct, European Commission, 12522/19 (2019).

2. Principles and objectives of the DSA

At the core of the DSA is the proper functioning of the internal market for online intermediary services.[94] The Act seeks to harmonise the regulations applying to intermediary services to ensure a safe and predictable online environment where fundamental rights are protected.[95] The DSA is also concerned with the exemption of intermediary service providers from liability and the due diligence obligations of such service providers.[96] The regulations introduce an exemption from liability even where intermediaries, in good faith, investigate and remove illegal content of their own volition.[97] The obligations of intermediary service providers under the DSA are tiered based on the size, nature and impact of the intermediary service provider.[98]

'Intermediary services' for purposes of the Act includes all conduit, caching and hosting services regardless of their place of establishment provided they have a substantial connection to the EU.[99] Additional obligations are provided for online platforms such as social networks as a subcategory of hosting services which not only store information for their users at their request but which also disseminate information to the public, that is make information easily accessible at the request of the recipients of the service.[100] The Act also places distinct and cumulative obligations on very large online platforms and search engines whose users make up to 10% of the EU population and which pose a great risk in the dissemination of illegal content and societal harm.[101]

3. Regulation of hate speech in the DSA

Under the DSA 'hate speech' is not defined and is classified as part of a general form of 'illegal content' which is unlawful under Union or national law.[102] What is illegal content is therefore dependent on the regional or national law considered.[103] As the DSA neither defines 'hate speech' nor repeals the

[94] DSA, Article 1 (1).
[95] *Ibid*.
[96] *Ibid*, Article 1 (2).
[97] *Ibid*, Article 7.
[98] *Ibid*, Recital 41.
[99] *Ibid*, Article 3(g) and (e); Directive EU 2015/1535 of the European Parliament and of the Council of 9 September 2015 laying down a procedure for the provision of information in the field of technical regulations and of rules on information society services (codification).
[100] DSA, Article 3 (i) (k)
[101] *Ibid*, Article 33 (2).
[102] *Ibid*, Recital 12.
[103] A. Manganelli and A. Nicita, *Regulating Digital Markets: The European Approach* (Palgrave Macmillan, 2022) at 192.

2008 Framework Decision, the definition of hate speech adopted in the Framework Decision remains valid despite its limitations to speech relating to 'race, colour, descent or national or ethnic origin.'[104]

Generally, all intermediary service providers are required to transparently state in their terms and conditions the restrictions applicable to their service, such as the kind of speech which is not permitted, and the way such restrictions will be enforced.[105] Additionally, intermediary service providers such as online platforms which provide a hosting service must, without undue delay, disable access to illegal content upon becoming aware of it whether by internal mechanisms or in pursuance of orders issued by national and administrative authorities.[106] Orders transmitted by national authorities to intermediary service providers must contain, among other things, a legal basis and statement of reasons explaining why the content is illegal.[107] Intermediaries are to inform the recipients of their services who are affected by such orders at the time stated by the issuing authority or, at the latest, upon giving effect to such an order.[108]

Providers of intermediary services may also conduct their own investigations into content and find it to be illegal or incompatible with their terms of service.[109] However, the DSA emphasises that there is no general monitoring or fact-finding obligation on intermediary service providers.[110]

In addition to the obligation to take-down or disable illegal content, the DSA places due diligence obligations on intermediary service providers. General due diligence obligations, such as the requirement to establish a point of contact, designate legal representatives and give notice of their terms and conditions, apply to all intermediary service providers.[111]

Providers of hosting services such as online platforms are required to create easily accessible notice-and-action mechanisms which allow any individual or entity to flag potentially illegal content.[112] The DSA assigns the role of "trusted flaggers" to entities with demonstrated expertise in countering illegal

[104] Council Framework Decision 2008/913/JHA of 28 November 2008 on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law, 2008 O.J (L 328) 55, Recital 9.
[105] DSA, Article 14 (1).
[106] *Ibid*, Article 9, Article 16 (3).
[107] *Ibid*, Article 9 (2) (i).
[108] *Ibid*, Article 9 (5).
[109] *Ibid*, Article 7.
[110] *Ibid*, Article 8.
[111] *Ibid*, Article 12, 13, 14.
[112] *Ibid*, Article 16.

content online.[113] Online platforms are required to give priority to complaints submitted by trusted flaggers to fast-track the procedure and increase accuracy.[114]

In as far as it creates procedural obligations in content moderation practices, the DSA arguably promotes the freedom of expression and protects the freedom of users to access online platforms without arbitrary exclusion.[115] However, in the context of regulating hate speech, the lack of a contemporary and uniform definition for 'hate speech' coupled with the obligation to comply with notice-and-takedown orders based on national laws does little to prevent censorship and/or harmonise the obligations of intermediaries.[116] Further, the almost judicial discretion granted to private companies in the moderation of content is dangerous to the freedom of expression as it may lead to censorship of legitimate speech or the proliferation of dangerous hate speech.[117] It is imperative that clear parameters be set for the nature of impugned content and the manner in which such content should be moderated especially with regard to hate speech.[118]

4. **Human rights and public policy obligations of intermediaries**

The content moderation procedures envisaged under the DSA raise two distinct human rights concerns: the application of human rights norms to orders issued by states to intermediaries regarding third party content and the human rights norms applicable to the content moderation activities of intermediaries which are privately owned companies and rights holders themselves. The former concern is addressed by the International Law Commission's Articles on State Responsibility which place human rights obligations on states for the actions of non-state actors which are attributable to states through instruction, direction or control.[119] The DSA attempts to address the latter concern by requiring all intermediaries to pay due regard to fundamental human rights in their content

[113] *Ibid*, Article 22 (1).

[114] *Ibid.*

[115] I. Tourkochoriti, "The Digital Services Act and the EU as the Global Regulator of the Internet" (2023) 24 *Chicago Journal of International Law* 129 at 135.

[116] D. K. Citron, "Extremist Speech, Compelled Conformity, and Censorship Creep", (2018) *Notre Dame Law Review,* 1050.

[117] Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, available https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf (date accessed: 1 August 2024).

[118] M. K. Land, "Against Privatised Censorship: Proposals for Responsible Delegation" (2020) 60 *Virgina Journal of International Law* 426; *Yildrim v. Turkey*, 2012 Eur. Ct. H. R. 505 at para. 64.

[119] United Nations General Assembly Res. 56/83, annex, Responsibilities of States for Internationally Wrongful Act (12 December 2001) Article 8.

moderation activities thus creating a horizontal effect of fundamental rights.[120] The Act highlights the importance of 'responsible and diligent behaviour' as essential for safety, trust and predictability online and key to the protection of the freedom of expression and information.[121] As with the obligations in relation to illegal content, the DSA addresses human rights and policy obligations of intermediaries asymmetrically.

All intermediary service providers are generally required to pay due regard to the relevant international standards for the protection of human rights.[122] The DSA commendably references the UNGP as an example of the international human rights standards which intermediary service providers should give due consideration.[123] However, this reference to international human rights law standards and the UNGP is to be found only in the recitals and is not repeated in the enacting provisions of the DSA. As discussed, the UNGP are instructive on the human rights considerations incumbent on business entities. There was therefore a missed opportunity to embed the principles of the UNGP into the substantive obligations of intermediaries under the DSA.

Additionally, all intermediary service providers are required to state clearly their terms of use and enforce these terms and conditions proportionately and with due regard for fundamental rights such as the freedom of expression.[124] However, the language of the DSA may foreseeably pose challenges to the interpretation of the scope of these obligations. Firstly, strictly reading Article 14 (4), the obligation to consider the fundamental rights of users only arises at enforcement and not at the creation of terms and conditions of use.[125] In contrast, recital 47 of the DSA clearly states that the fundamental rights of users must be considered in 'designing, applying and enforcing' restrictions. Whether the full scope of obligations described in the recitals can be read into the enacting provisions is subject to interpretation. Secondly, it is unclear whether 'proportionality' is used in reference to the principle in international law which guides state action.[126] If this is the case, a secondary issue that may arise for intermediaries is how to navigate the ambivalence between regional and international

[120] W. Schulz and C. Ollig, "Hybrid Speech Governance: New Approaches to Govern Social Media Platforms under the European Digital Services Act?" (2023) 14 *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* at 575.
[121] DSA, Recital 3.
[122] *Ibid*, Recital 47.
[123] *Ibid*, Recital 47.
[124] DSA, Article 14 (4).
[125] Schulz and Ollig, *supra* note 120 at 576.
[126] *Ibid,* at 574.

standards of human rights law regarding the test of proportionality in determining how hate speech should be treated as discussed in the preceding section of this paper.[127]

Intermediary service providers which offer hosting services are particularly required to observe the fundamental rights of the recipients of their services when removing or disabling access to illegal content.[128] Online platforms such as social media sites are required to give due consideration to fundamental rights in actions following a notice of illegal content, particularly, they are to only remove or disable access to specific items or information which constitute illegal content without unduly affecting the freedom of expression.[129]

In addition to a requirement to pay due regard to the freedom of expression in designing and applying restrictions to speech,[130] very large online platforms are required to conduct risk assessments of the foreseeable or actual impact of their services on the enjoyment of fundamental rights, including the freedom of expression, arising from factors such as the design of the algorithms used by such platforms.[131] These very large platforms are then required to proactively mitigate the risks identified in a manner that is both proportionate to their economic capacity and gives particular consideration to the freedom of expression.[132]

Mitigation measures include the adaption of content moderation systems, algorithmic systems and internal processes, capacity building and development of local expertise.[133] The mitigation measures adopted are to be 'reasonable, proportionate and effective', giving special consideration to the freedom of expression.[134] Although the protection of freedom of expression is particularly referenced, it is once again unclear what proportionality means in the context of intermediaries as private actors. As the Commission is yet to issue guidelines on the application of these mitigation measures, it remains to be seen how these mitigation measures will be balanced against the freedom of expression.[135]

The application of human rights standards to the regulation of hate speech by intermediary service providers in pursuance of the DSA may not be as straightforward as the Act suggests. While "due regard" must be given to fundamental rights, no substantive scope is given leaving the extent of the

[127] DSA, Article 14 (4).
[128] *Ibid*, Recital 22.
[129] *Ibid*, Recital 51.
[130] *Ibid*, Recital 47.
[131] *Ibid*, Article 34.
[132] *Ibid*, Recital 86.
[133] DSA, Article 35.
[134] *Ibid*, Article 35 (1).
[135] *Ibid*, Article 35 (3).

obligation open to interpretation. A foreseeable challenge to upholding the freedom of expression in the context of designing restrictions to curb hate speech is the lack of consensus on the definition of 'hate speech' and the limited range of protected characteristics in EU law. As discussed, the regional jurisprudence of the ECtHR on hate speech departs from international standards in several material aspects.[136] If the UN standard of protection of the freedom of expression is to be applied to the moderation of hate speech, online platforms will have to apply principles such as the least intrusive measure and a test for proportionality as prescribed by the Human Rights Committee.[137]

Content moderation by intermediaries at present resembles automated routine industrial processes rather than a nuanced process applying international human rights law standards on a case-by-case basis.[138] The impact of the DSAs attempts to apply human rights norms and public policy obligations in the regulation of hate speech by intermediaries is likely to be limited to *ex post* considerations of the appropriateness of decisions taken by intermediaries rather than *ex ante* applications of human rights norms in content moderation.[139] Consequently, without judicial or legislative interpretation of the human rights obligations of intermediaries in content moderation, the DSA approach is likely to have a limited impact in the ongoing efforts to develop substantive human rights obligations for intermediaries.

In proposing an "international law of the internet" Land argues that Article 19 of the ICCPR is broad enough to place direct human rights obligations on intermediaries whose activities have an impact on public discourse.[140] She argues convincingly in favour of imposing a limited range of state-like human rights obligations on intermediaries which assume a dominant market position with regard to expressive rights and therefore threaten the freedom of expression.[141] Such intermediaries ought to be subject to the same legality, legitimacy, proportionality and due process requirements as states in their content moderation practices.[142] While it is doubtful that the DSA in its current construction can be interpreted to impose such direct obligations, it is certainly a step in the direction of establishing

[136] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 26-27, U.N. Doc. A/74/486 (Oct. 9, 2019).
[137] Aswad and Kaye, *supra* note 28 at 211.
[138] P. Leerssen, "An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation" (2023) 48 *Computer Law & Security Review* 1 at 5.
[139] *Ibid.* at 7.
[140] M. K. Land, "Toward an International Law of the Internet" (2013) 54 *Harvard International Law Journal* 393 at 445.
[141] *Ibid.* at 447.
[142] *Ibid.*

the human rights responsibilities of intermediaries which will likely influence similar action from law makers in other jurisdictions.[143]

5. **Accountability and transparency in content moderation in the DSA**

The DSA's regulation of the internal moderation procedures of intermediaries conducted based on their terms and conditions is a novelty, effectively requiring accountability in the enforcement of these previously unregulated terms and conditions.[144] The DSA also establishes independent administrative authorities at national and regional level such as the Digital Service Coordinator and the European Board for Digital Services which will oversee compliance with the DSA in content moderation.[145]

The transparency obligations in content moderation introduced by the DSA will be analysed with reference to three forms of transparency proposed by Machado and Aguiar: broad transparency including through reporting obligations; systemic transparency which relates to algorithmic transparency; and individual transparency relating to notice to the user about how content is moderated, and decisions are made.[146]

With regard to broad transparency, under the DSA online platforms are required to submit yearly reports on their content moderation activities whether conducted of their own volition or on receipt of an order from a national authority.[147] Additionally, providers of online platforms are to submit their decisions and statements of reasons in relation to their content moderation activities.[148] Very large online platforms are required to make publicly available annual reports on their content moderation including measures taken in enforcement of their terms and conditions.[149] They are also required to provide access to their data to allow effective assessment of their compliance and are to be subjected to independent audits.[150]

As regards systemic transparency, very large online platforms are required to conduct risk assessments of the foreseeable or actual impact of their services including, among other things, on the enjoyment of the freedom of expression, arising from factors such as the design of the algorithms used by such large online platforms.[151] They are also to, where possible, voluntarily provide publicly

[143] A. Bradford, "The Brussels Effect" (2012) 107 *Northwestern University Law Review* 1.
[144] *Ibid.* at 5.
[145] DSA, Chapter IV.
[146] Machado and Aguiar, *supra* note 60 at 251.
[147] DSA, Article 15 (1).
[148] *Ibid*, Article 24 (5).
[149] *Ibid*, Recital 49.
[150] *Ibid*, Article 37, 40.
[151] *Ibid*, Article 34

accessible data such as interactions and engagement to stakeholders such as researchers who are concerned with monitoring societal concerns.[152]

With regard to individual transparency, the DSA requires that where a hosting service such as an online platform determines that content provided by a user is illegal or incompatible with its terms of service and removes or disables access to it, they must provide a clear and comprehensible statement of reasons to the affected user and inform them of the avenues for redress available to them.[153] Users of intermediary services ought to be able to easily and effectively contest content moderation decisions of online platforms through internal complaint-handling mechanisms which allow users to complain without strict formality and external dispute resolution.[154]

The transparency and accountability provisions of the DSA set out key principles of due process which address many of the concerns raised over the lack of oversight over content moderation by intermediaries.[155] Particularly it establishes reporting obligations, sets minimum standards for content moderation and appeal, establishes oversight mechanisms and creates systems for regular audits which were previously not part of regulatory regimes for intermediary service providers.[156] Schulz and Ollig argue that these procedural safeguards could promote legal certainty in the substantive human rights obligations of intermediaries under the DSA.[157] However, they concede that without a binding determination or legislative clarification, the substantive human rights obligations for intermediaries particularly relating to the freedom of expression will remain undefined.[158]

6. **Conclusion**

The DSA is an important step forward for the creation procedural obligations in content moderation practices and the promotion of the freedom of expression online without arbitrary exclusion.[159] However, in the context of regulating hate speech, the lack of a contemporary and uniform definition for 'hate speech' coupled with the obligations to comply orders based on national laws potentially undermines this. Without clear parameters for the nature of impugned content and oversight over the manner in which such content should be moderated especially with regard to hate speech, there is potential for legitimate speech to be stifled.

---

[152] *Ibid*, Recital 98

[153] DSA, Article 17.

[154] *Ibid*, Article 20, 21.

[155] Gillespie, *supra* note 57 at 213.

[156] *Ibid.* at 214 – 215.

[157] Schulz and Ollig, *supra* note 120 at 577.

[158] *Ibid.*

[159] I. Tourkochoriti, "The Digital Services Act and the EU as the Global Regulator of the Internet" (2023) 24 *Chicago Journal of International Law* 129 at 135.

While DSA commendably references the UNGP as an example of the international human rights standards which intermediary service providers should give due consideration, this reference is only in the recitals and is not repeated in the enacting provisions of the DSA. Potential challenges arise from the interpretation of 'hate speech' to different standards of protection across regional human rights systems and the extent of the human rights obligations of intermediaries with the impact that the approach of the DSA will only be effective for providing considerations of the appropriateness of decisions taken by intermediaries after the fact rather than the applications of human rights norms at all stages from the development to implementation of speech restrictions.

Finally, the DSA introduces accountability and transparency mechanisms to the previously unlegislated internal moderation procedures of intermediaries based on their terms and conditions. The transparency mechanisms adopted under the DSA are broad, systemic and individual, incorporating key principles of oversight and due process. However, without a binding determination or legislative clarification on the substantive human rights obligations of intermediaries, the effectiveness of these accountability and transparency mechanisms in ensuring that the freedom of expression is upheld will be severely limited.

D. CONTEMPORARY APPROACHES TO REGULATING HATE SPEECH ONLINE

1. Intermediaries as regulators: platform laws and commercial content moderation

The EU approach to regulating intermediaries attempts to place accountability, transparency and human rights obligations on intermediaries putting them in the position of co-regulators which actively make decisions on the acceptable limits of the freedom of expression online while simultaneously being outside the direct reach of international human rights law.[160] This model grants authority to intermediaries to "translate legal principles established for offline content into the online environment."[161] As profit driven entities, intermediaries are not primarily concerned with the overall impact of their decisions on the exercise of fundamental freedoms and must balance their commercial interests with these growing obligations.[162]

[160] F. R. Jørgensen and A. M. Pedersen, "Online service providers as human rights arbiters" in M. Taddeo and L. Floridi (eds.) *The Responsibilities of Online Service Providers* (Oxford: Oxford University Press) at 179.
[161] Land, *supra* note 140 at 406.
[162] Gillespie, *supra* note 57 at 199.

Through their terms and conditions and community guidelines or "platform laws",[163] intermediaries determine the acceptable limits of speech, create rules on how content deemed hate speech will be handled and enforce those rules.[164] Intermediaries generally have wide discretion in determining their terms and conditions of use and 'community standards' which are largely influenced by their commercial interests.[165] As stated by Gillespie, content moderation is the "central value proposition" of intermediaries.[166] The DSA is a laudable attempt to impose transparency and accountability obligations on intermediaries in the enforcement of their terms and conditions, however, it does not address the lack of transparency in the processes by which these guidelines and policies are created.

Content moderation happens at a massive scale through routine processes driven by algorithms and user flagging.[167] Most intermediaries employ a complex moderation structure combining algorithms, reporting procedures and commercial content moderators to detect and review hate speech online.[168] Moderation can be *ex ante*, prior to publishing of content or *ex post*, after content is published.[169] *Ex ante* regulation is characterised by algorithmic regulation and involves the encoding of legal or community values into platform software code, transformation of input data, decision-making and regulatory governance processes which are adaptive in nature and employ machine learning to screen content before it is shared.[170] The design of algorithms therefore greatly impacts the manner in which content is moderated online.[171] Examples of *ex ante* moderation are hashing which identifies illegal material based on unique signatures and geo-blocking which limits access to content within particular locations usually at the behest of a government.[172] As with traditional paternalistic conceptions of

[163] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/38/35 (6 April 2018).

[164] R. A. Wilson and M. K. Land, "Hate speech on Social Media: Content Moderation in Context" (2021) 52 *Connecticut Law Review* 1029 at 1045..

[165] *Ibid*.

[166] Gillespie, *supra* note 57 at 201.

[167] T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (London: Yale University Press, 2018) at 23.

[168] C. Katzenbach and L. Ulbricht, "Algorithmic governance: Internet Policy Review" (2019) 8 *Alexander von Humboldt Institute for Internet and Society* 1 at 10.

[169] K. Klonick, "The New Governors: The People, Rules and Processes Governing Online Speech" (2018) 131 *Harvard Law Review* 1598 at 1635.

[170] K. Yeung, "Algorithmic regulation: A critical interrogation" (2018) 12 *Regulation and Governance* 505 at 507.

[171] A. Manganelli and A. Nicita, *Regulating Digital Markets: The European Approach* (Palgrave Macmillan, 2022) at 192.

[172] K. Klonick, "The New Governors: The People, Rules and Processes Governing Online Speech" (2018) 131 *Harvard Law Review* 1598 at 1637.

regulation by code, algorithmic regulation depends on intermediaries as regulators to set the standard of acceptable conduct.[173]

However, despite their complexity, algorithms cannot detect cultural and contextual nuance of speech and are therefore not completely accurate.[174] This may disproportionately affect marginalised communities and minorities and lead to the spread of hate speech.[175] Furthermore, algorithms can be influenced by discriminatory assumptions and biases built into code which are difficult to detect.[176] Additionally, where machine-learning is employed in algorithmic design, the complete decision making processes and their outcomes are unknown even to the designers of such algorithms as the process continues to develop after the original programming is completed.[177]

Ex post content moderation takes place manually either on a proactive basis, by intermediaries of their own initiative, or reactively through flagging by users which is then reviewed by human moderators.[178] Intermediaries employ a large human workforce of commercial content moderators to conduct content evaluation in addition to algorithmic regulation processes. However, these moderators are typically limited to deciding on moderation actions through snap judgements based on internal guidelines without sufficient time to carefully deliberate or investigate the impugned speech.[179] The guidelines which these content moderators apply are not often publicly accessible and change frequently.[180] Further, moderators are seldom well-informed on different local contexts and do not have the linguistic expertise to effectively moderate content from all the markets in which these intermediaries operate.[181]

Commercial content moderation also creates additional problems beyond the freedom of expression. Given the sheer volume of content which intermediaries screen, an ecosystem of labour has

[173] Murray, *supra* note 54 at 81.

[174] Tourkochoriti, *supra* note 115 at 141; Council of Europe, Committee of experts on internet intermediaries, Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications, available https://rm.coe.int/study-on-algorithmes-final-version/1680770cbc at 21 (date accessed: 11 July 2024).

[175] Sander, *supra* note 61 at 957.

[176] L. McGregor, D. Murray and V. Ng, "International Human Rights Law as a Framework for Algorithmic Accountability" (2019) 68 *International & Comparative Law Quarterly* 309 at 317.

[177] *Ibid* at 319.

[178] Klonick, *supra* note 172 at 1638.

[179] S. T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (New Haven: Yale University Press, 2019) at 34.

[180] C. Buni and S. Chemaly, "The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech", *The Verge* (13 April 2016).

[181] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/38/35 (6 April 2018) at 11.

developed to meet the ever-growing demand and maintain the internal functioning of online platforms and other intermediaries.[182] In efforts to reduce the cost of content moderation, intermediaries have outsourced content moderation jobs to third-parties where "ghost workers" are employed at exploitative rates.[183] These moderators also suffer psychological effects from exposure to volumes of abhorrent content.[184]

In 2022, the European Commission released its results of the seventh evaluation of the 2016 Code of Conduct.[185] Its findings were that while 71% of signatories removed flagged content within 24 hours of notice, flagged content was not evaluated consistently over time. Further, intermediaries fell short in ensuring transparency and proving feedback to users. Additionally, only Facebook informed its users systematically where their content was subject to sanctions. The periodic evaluations and reports by the signatories to the 2016 Code of Conduct do not, however, provide much detail of the commercial content moderation processes undertaken to achieve the results reported.[186] Nor do they indicate the limits to algorithmic and human moderation processes, including where such processes proved inadequate in identifying and addressing hate speech.

2. Challenges to regulating hate speech online at a regional level

The regulation of hate speech in the online public sphere presents challenges that are both substantive and procedural in nature. This is further complicated when one considers the cultural and legal differences between states and across regions. This section discusses some of the challenges which a regional approach to regulating speech such as the DSA is likely to face.

Firstly, determining whether speech falls within the ambit of 'hate speech' requires an assessment of the social, cultural and relational context in which it occurred.[187] Liberal democratic values may render speech allowable as a valid exercise of the freedom of expression in some jurisdictions and illegal in others. Even where there is general agreement on the harms of speech, such as the European position

[182] Roberts *supra* note 179 at 203.

[183] Mary L. Gray and Siddarth Suri, *Ghost Work: How to Stop Silicone Valley from Building a New Global Underclass* (Boston: Houghton Mifflin Harcourt, 2019) at 38.

[184] Roberts, *supra* note 179 at 209.

[185] European Commission, *7th evaluation of the Code of Conduct* (November, 2022) available https://commission.europa.eu/ (date accessed: 11 July 2024).

[186] Information provided by the IT companies about measures taken to counter hate speech, including their actions to automatically detect content, available https://commission.europa.eu/ (date accessed: 11 July 2024).

[187] B. Farrand, "'Is This a Hate Speech?' The Difficulty in Combating Radicalisation in Coded Communications on Social Media Platforms" (2023) 29 *European Journal on Criminal Policy and Research* 477at 479.

informed by the history of hate speech preceding the gravest violations of human rights in the Holocaust, there is no consensus among states on how Holocaust denial ought to be treated legally.[188] In contrast, in the US hate speech is generally protected except in the case of imminent violence.[189]

Furthermore, research has shown that hate speech online can be incredibly nuanced and coded, framed as humorous through non-textual mixed media such as memes and terms which would otherwise not be considered hateful or derogatory.[190] Humour is increasingly being used by extremists as a means of "blurring the lines between mischief and potentially radicalising messages."[191] Speech which is presented as humorous or ironic creates a 'hate-humour nexus' potentially widening the scope of allowable speech in the direction of hateful rhetoric.[192] The re-adoption and changed meaning of words by marginalised communities in attempts to reclaim words intended to incite hate may also be difficult to understand and account for by persons not familiar with a counter culture.[193] This poses an additional challenge for both human and algorithmic regulators in identifying hate speech in subtext.[194]

As a result, training both algorithms and human moderators to accurately identify prohibited content while giving due consideration to the nuance inherent in speech from different cultural contexts is a difficult task.[195] In the context of regulating hate speech, algorithms developed primarily in Western countries based on human knowledge limited to that context will be unable to identify nuance in speech from other parts of the world.[196] This could potentially lead to content from specific geographical communities being disproportionately subjected to sanctions.[197]

---

[188] J. Menkes, "Freedom of speech in the age of digitalization: Opportunities and threats" in L. D. Dabrowski and M. Suska (eds) *The European Union Digital Single Market: Europe's Digital Transformation* (Routledge, 2022).
[189] *Brandenburg v Ohio*, 395 U.S 444 (1969).
[190] Farrand, *supra* note 187 at 480.
[190] J. C. York and E. Zuckerman, "Moderating the Public Sphere" in F. R. Jørgensen (ed.) *Human rights in the Age of platforms* (MIT Press, 2019) at 152.
[191] M. Fielitz and R. Ahmed, "It's not funny anymore: Far rights extremists' use of humour", European Commission, Radicalisation Awareness Network (2019).
[192] T. Askanius, "On frogs, monkeys, and execution memes: Exploring the humour-hate nexus at the intersection of neo-Nazi and alt-right movements in Sweden" (2021) 22 *Television & New Media* 147 at 152.
[193] C. Peters and S. Allan, "Weaponising memes: The journalistic mediation of visual politicisation" (2022) 10 *Digital Journalism* 217.
[194] Farrand, *supra* note 187 at 481; P. Bhat and O. Klein "Covert hate speech: White nationalists and dog whistle communication on Twitter" in G. Bouvier and J. E. Rosenbaum (eds.) *Twitter, the Public Sphere, and the Chaos of Online Deliberation* (Springer, 2020) 151 - 172.
[195] York and Zuckerman, *supra* note 190 at 152.
[196] K. Crawford, "Artificial Intelligence's White Guy Problem" *New York Times* (25 June 2016).
[197] G. Schlag, 'European Union's Regulating of Social Media: A Discourse Analysis of the Digital Services Act' (2023) 11 *Politics and Governance* 168 at 170.

Regarding the procedural challenges to regulation, a flagging-based model presents further concerns. The notices in the notice-and-takedown model form the beginning of what Roberts refers to as the "commercial content moderation cycle of review" and are an integral part of the speech regulation performed by intermediaries.[198] The quality of these notices and the technical procedures for their submission all have an impact on the manner and speed with which notices are processed.[199] Notably, the DSA attempts to demystify the process by setting procedural guidelines for intermediaries, especially online platforms, and maintaining the role of trusted flaggers.[200] However, in order for this model to be effective, a high level of technical and contextual knowledge on the part of trusted flaggers is needed.

Additionally, notice-and-take down procedures which rely on flagging of content may potentially be abused to stifle opposing views.[201] Content moderation and flagging processes are subject to "system designs, multiple actors and intentions, assertions and emotions."[202] As a result, these procedures are in danger of being used for censorship and media manipulation.[203] Control over what is allowable speech through illegitimate flagging may lead to the creation of echo chambers which do not allow diverse views and discussion to flow freely.[204]

Finally, enforcement against intermediaries is likely to present some challenges particularly where wording such as "have due regard" is used in reference to the freedom of expression. The DSA is ambitious in its attempts to simultaneously underscore immunity from liability and place enforceable obligations against intermediaries. It remains to be seen how these obligations will be enforced, if at all, given the differences in opinion even within the EU as demonstrated by the Republic of Poland's challenge to the imposition of obligations on online platforms to make efforts to ensure the unavailability of certain works.[205]

---

[198] Roberts, *supra* note 179 at 220.

[199] K. Crawford and T. Gillespie, "What is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint" (2014) 18 *New Media & Society* 420.

[200] DSA, Article 22 (1).

[201] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/32/38 (11 May 2016) at 14.

[202] Crawford and Gillespie, *supra* note 199 at 411.

[203] *Ibid* at 420; N. Strossen, *HATE: Why We Should Resist It with Free Speech, Not Censorship* (Oxford University Press, 2018).

[204] E.B. Laidlaw, "A Framework for Identifying Internet Information Gatekeepers" (2010) 24 *International Review of Law, Computers and Technology* 263.

[205] *Republic of Poland v. European Parliament and Council of the European Union,* Case C-401/19 (2022).

3. *Ubuntu* and communitarianism in combatting hate speech: distilling an African regional approach

Ubuntu is a Nguni word which has become synonymous with the African values of human dignity and communitarianism.[206] The term derives from the phrase '*umuntu ngumuntu ngabantu*' translated as 'a person is a person through other persons' and forms the basis of the African moral-philosophical principles on which an array of individual and collective rights are derived.[207] *Ubuntu* is both metaphor for social, legal and ethical judgment of conduct as well as a barometer for measuring the propriety of actions.[208] The interests of the people, both the individual and the community, must form the primary concern of individual, community and state action.[209] These communitarian values are exhibited in a range of principles that are concomitant with what are today described as human rights and democratic values such as unconditional human dignity, participatory decision-making, respect for the inherent capacity for people to commune with each other and mutual recognition of diversity.

Ubuntu does not, contrary to some arguments, posit that preserving community must supersede all other concerns to the detriment of the individual,[210] rather it upholds individuation by requiring "tolerance, understanding and respect towards all individuals in interpersonal relationships, in relations between the individual and the groups of which she forms part, between different groups… between different communities."[211] Furthermore, to describe *ubuntu* as an African philosophy does not suggest that African cultures are a monolith, rather that in the wide cultural diversity and across different indigenous African traditions communitarian characteristics are distinguishable.[212]

Ubuntu is a uniquely African perspective which offers potential for solutions to uniquely African problems including for the control of hate speech online and the regulation of intermediaries.[213] An

[206] M. B. Ramose, "The philosophy of ubuntu and ubuntu as philosophy" in P. H. Coetzee and A. P. J. Roux (eds), *Philosophy from Africa: A text with readings* (Oxford University Press, 2002) at 30; T. Metz, "African Values and Human Rights as Two Sides of the Same Coin: A Reply to Oyowe" (2014) 14 *African Human Rights Law Journal* 306 at 308.

[207] T. Metz, "African Values and Human Rights as Two Sides of the Same Coin: A Reply to Oyowe" (2014) 14 *African Human Rights Law Journal* 306 at 308.

[208] D. D. Ndima, "Reconceiving African Jurisprudence in a Post-Imperial Society: The Role of Ubuntu in Constitutional Adjudication" (2015) 48 *Comparative and International Law Journal of South Africa* 359 at 370.

[209] *Ibid*.

[210] A. Oyowe, "Strange bedfellows: Rethinking *ubuntu* and human rights in South Africa" (2013) 12 *African Human Rights Law Journal* 103.

[211] M. Pieterse, "Traditional African Jurisprudence" in C. Roederer and D. Moellendorf (eds.) *Jurisprudence* (Juta Lansdowne, 2004) at 445.

[212] K. Wiredu, "Social philosophy in postcolonial Africa: Some preliminaries concerning communalism and communitarianism" (2008) 27 *South African Journal of Philosophy* 332.

[213] Pieterse, *supra* note 211 at 438.

approach to the regulation of hate speech online in the African context which applies the principles of *ubuntu* is one which is human centred rather than market driven.[214] It is characterised by participation of communities in decision-making processes especially at the stage of setting limits to speech through the development of terms and conditions and community guidelines. Rather than the one-size fits all approach that has so far characterised intermediary policies, a clear representation of community values and contextually relevant indigenous knowledge will be indispensable to the effective regulation of hate speech online.

Thus, an approach to the regulation of hate speech which has as its basis indigenous African values of communitarianism departs from the western goal of protecting the individual as a consumer and rather seeks to protect a community or people and their ability to relate communally.[215] While, as argued by Monsees and Lambach, western regulation is characterised by an undercurrent of global competition over technology, business and infrastructure,[216] an African approach will be anchored in preserving relationships between the individual as a rights holder and the community and between communities or peoples as rights holders through open discourse.

An approach to intermediary regulation centred on participatory decision making and recognition of indigenous knowledge systems such as *ubuntu* is not an entirely novel proposition. DeBurca for instance proposes a model of global experimentalist governance which involves an institutionalized transnational participatory model of regulation wherein issues are framed "in an open-ended way and subject to periodic revision by various forms of peer review in the light of locally generated knowledge.[217] Although aware of the goals they wish to attain, experimentalists understand the limitations to their ambitions and therefore revise their procedures on the basis of experience and establish accountability mechanisms.[218] Similarly, the model of multistakeholderism first proposed by the Working Group on Internet Governance and since endorsed by several scholars emphasises participation by various stakeholders in the creation of norms, rules, decision-making procedures and programs that shape the evolution and use of the internet.[219]

---

[214] Metz, *supra* note 207 at 308.

[215] *Ibid* at 318.

[216] L. Monsees and D. Lambach, "Digital sovereignty, geopolitical imaginaries, and the reproduction of European identity" (2022) 31 *European Security* 377 at 379.

[217] G. De Burca, R. O. Keohane and C. F. Sabel, "Global Experimentalist Governance" (2014) 44 *British Journal of Political Science* 477 at 479.

[218] *Ibid* at 483.

[219] United Nations Working Group on Internet Governance, 3 August 2005, WSIS-II/PC-3/DOC/5-E, at para. 10; R. H. Weber, 'A Legal Lens into Internet Governance' in Laura DeNardis et al (eds) *Researching Internet Governance: Methods, frameworks, futures* (Cambridge Massachusetts: 2020 MIT Press, 2020); R.H Weber, "Elements of a Legal Framework for Cyberspace" (2016) 26 *Swiss Review of International and European Law* 195.

4. Recommendations for regulating online hate speech in the African context

To address the issues arising from the control of hate speech by intermediaries – lack of a common standard for identifying hate speech, lack of clear human rights obligations for intermediaries and transparency and accountability in content moderation processes – a voluntary code of conduct or non-binding guiding principles for online intermediaries are proposed as the ideal framework for the African context. Unlike the EU, the AU does not operate on a pooling of sovereignty and therefore does not exercise legislative power.[220] As such, an Act such as the DSA would find no translation in the African regional context. However, the ACHPR is mandated to formulate principles and rules aimed at solving legal problems relating to the enjoyment of fundamental rights on which African governments may base their legislation.[221] If framed as a regional human rights effort, the ACHPR may take a leadership role in negotiating the terms of such a framework as well as mobilising the requisite political will. This approach will also enhance the neutrality of such an intervention by freeing it from overbearing governments which may seek to hijack the process to advance ulterior motives.

The DSA is a culmination of decades long efforts characterised by consensus building and negotiations. However, it does little to change the status quo of obligations of intermediaries in relation to the regulation of hate speech from the position of the 2016 Code of Conduct. As such, while the scope of the international human rights law obligations of intermediaries develops, a non-binding framework which emphasises principles of *ubuntu* concurrently with the obligations under the UNGP and is similar to the 2016 Code of Conduct which many intermediaries have already voluntarily agreed to, will likely enjoy more buy-in as a step towards intermediary regulation which is appropriate for the African context. In the alternative, it is proposed that rather than adopt a single framework for the African continent, given the varying environments and historical contexts, different non-binding frameworks may be adopted in the 8 sub-regional communities recognised by the AU. This would enhance the contextual relevance of such a framework and allow intermediaries to adapt their practices to meet the highest standard in each environment. However, it would slow progression towards regional consensus on best practices.

Regarding the first issue of defining hate speech, a framework adopted to regulate such speech ought to require that platform terms of service be precise and accessible to users. While the national governments of individual states may legislate against hate speech and similar online harms, it is unlikely that all intermediaries will be aware of the parameters of such laws except in jurisdictions

[220] Treaty on the Functioning of the European Union, Official Journal of the European Union C 326/47, Article 114.

[221] African Charter, Article 45 (1) (b).

where they consider themselves particularly exposed such as where they have a large market share.[222] To avoid uncertainty in the applicable definition of hate speech and limits of legitimate speech, intermediaries ought to apply the international standard for the limitation of the freedom of expression as defined by the Human Rights Committee and adopted by the ACHPR in their content moderation activities on the continent.

The framework must place duties on intermediaries to conduct content moderation on a human rights rather than a risk averse basis in line with the UNGP. Kaye rightly proposes that content moderation activities of online platforms be subject to the tests of legality, legitimacy and necessity.[223] The question of legality can be answered by requiring, in similar terms as the DSA, that intermediaries set out in their terms of use including what kind of content violates these terms in clear unambiguous language.[224] As regards legitimacy, a wider range of protected characteristics ought to be reflected in an African regional approach. While the ICCPR and CERD ought to be instructive, the possible range of characteristics which may be subject to hate speech have grown since the adoption of these instruments. Restrictions to speech must therefore include a wider contextually informed range of protected characteristics.

To meet the necessity requirement, platforms ought to apply the least intrusive and proportionality tests in enforcing their terms and conditions as prescribed by the Human Rights Committee.[225] Intermediaries ought to adopt diverse restrictive measures as sanctions for violation of terms and conditions. Applying the test of proportionality in regulating hate speech online requires that intermediaries commit to considering the linguistic and local socio-political context in which they operate and in which speech is shared.[226] Particularly with regard to hate speech, intermediaries ought to consider tensions between communities as well as "the tone and content of the speech, the person inciting hatred, and the means of disseminating the expression of hate."[227] The legality, legitimacy and necessity tests ought to apply to formal and informal attempts by states to control the content

[222] C. Reed, "Command and Control" in Chris Reed, *Making Laws for Cyberspace* (Oxford University Press, 2012) 14.

[223] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, UN Doc A/HRC/38/25, 6 April 2018, para. 45.

[224] *Ibid,* para. 46.

[225] United Nations Human Rights Committee, General Comment No. 34, U.N. Doc. CCPR/C/GC/34 (Sept. 12, 2011), at para. 34.

[226] Sander, *supra* note 61 at 979.

[227] Report of the Special Rapporteur on the Promotion & Protection of the Right to Freedom of Opinion & Expression on Its Sixty-Seventh Session, UN Doc A/67/357, at para. 46; United Nations High Commissioner for Human Rights, Report on the Expert Workshop on the Prohibition of Incitement to National, Racial or Religious Hatred A/HRC/22/17Add.4 (Rabat Plan of Action on the Prohibition of Incitement to National, Racial or Religious Hatred that Constitutes Incitement to Discrimination, Hostility or Violance) at para. 29.

moderation process. The Global Network Initiative's Principles on Freedom of Expression and Privacy Implementation Guidelines are instructive on how intermediaries can honour human rights standards when faced with pressure from states.[228]

The development of user terms, moderation guidelines and algorithms which impact the freedom of expression online as well as the enforcement of such terms must also embody *ubuntu.* In that regard, purposeful engagement with local stakeholders and communities should form the key considerations both in the conception of these user terms, internal guidelines and algorithms, and in determining their linguistic and cultural appropriateness for the regional context.[229] Such engagement ought not to be limited to preliminary considerations but must be periodic and deliberate.[230] Engagement ought also to precede changes in user terms and internal moderator guidelines which may impact the freedom of expression.[231] This will address the issue of nuance in online hate speech and the need for a tailored approach to tackling hate speech.

As the principles elaborated in the UNGP do not distinguish between businesses based on their size, a framework to tackle hate speech in the African region ought similarly to place uniform requirements on intermediaries. [232] Due consideration ought to be given, however, to the size, market share and nature of platforms in determining their standard of compliance with the framework. Sander for instance states that larger platforms ought to exercise particular vigilance, especially in tense or conflict-ridden contexts, and sustain engagement with local stakeholders in "identifying prohibited forms of hate speech…and ensuring its timely removal in order to protect individuals and communities that may be adversely impacted by such speech."[233]

In addition to urging intermediaries to adopt human rights standards in the creation of platform terms and conditions, a framework for the African context on the regulation of hate speech online ought to address the application of international human rights law standards in the content moderation process. Intermediaries ought to commit to sustained human rights due diligence characterised by independent human rights audits and impact assessments in respect of their content moderation

[228] Global Network Initiative, Principles on Freedom of Expression and Privacy, available https://globalnetworkinitiative.org/wp-content/uploads/2018/04/GNI-Principles-on-Freedom-of-Expression-and-Privacy.pdf (date accessed: 21 July 2024).

[229] Access Now, Protecting Free Expression in the Era of Online Content Moderation (May 2019), available https://www.accessnow.org/wp-content/uploads/2019/05/AccessNow-Preliminary-Recommendations-On-Content-Moderation-and-Facebooks-Planned-Oversight-Board.pdf (date accessed: 30 June 2024).

[230] UNGP, Principles 17-19.

[231] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/38/35 (6 April 2018) at para. 55.

[232] UNGP, principle 14.

[233] Sander, *supra* note 61 at 983.

processes, especially those which rely on algorithms.[234] Due diligence processes ought necessarily to include assessments to identify actual and potential adverse impacts of intermediary activities on human rights, mitigation initiatives and transparent communication of the findings of these processes to users.[235]

Drawing from the approach of the DSA, the framework envisioned ought to encourage transparency and accountability through publicly accessible reports on the processes leading to content moderation including through algorithmic regulation and distinguishing between actions pursued in pursuance of court orders, governmental directives or user-generated notices. Such transparency should be both qualitative and quantitative providing stakeholders with information not only of actions taken but of the processes underlying such actions including flagging procedures and the internal guidelines which dictate these processes.[236] This ought to include data regarding the human content moderators engaged, the accuracy of their work and their working conditions addressing both wages and psychological and emotional well-being.[237]

The framework envisioned here would employ the notice-and-action approach of the 2016 Code of Conduct. As flagging mechanisms depend on quality notices and institutions designated as trusted flaggers, the framework ought to make provision for intermediaries to engage particularly with local civil society actors who have contextually relevant information including through capacity building to enable them to perform effectively as flaggers. Quantitative and qualitative data on content moderation processes ought to be accessible to such civil society actors and other stakeholders to develop industry-wide best practices and consistent metrics for measuring intermediary accountability.[238]

The costs of setting up and complying with the framework proposed here, as well as building institutional capacity would be borne primarily by intermediaries as part of their corporate social responsibility. Under the DSA, a supervisory fee is charged on very large online platforms and search engines to meet similar costs.[239] This fee is proportionate to the size of the online platform or online search engine and the number of people receiving their services within the region. A similar but considerably less onerous financial burden in the case of the framework proposed here is therefore

[234] UNGP, principle 17; McGregor, Murray and Ng, *supra* note 176 at 330.
[235] UNGP, principle 21.
[236] UNGP, principle 20 (a).
[237] Roberts, *supra* note 179 at 210; Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/38/35 (6 April 2018) at para. 57.
[238] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/38/35 (6 April 2018) at para. 52.
[239] DSA, preamble, para. 101.

justifiable given that large platforms obtain economic value from the data obtained from users through largely extractivist practices particularly in the Global South.[240] A pledge to direct some of these funds into ensuring that users rights are respected is not only reasonable but in line with the duty of business entities to take adequate measures to prevent, mitigate and remedy the adverse human rights impacts of their businesses under the UNGP.[241]

While the framework envisioned in this work seeks to ensure that content moderation is grounded on a human rights approach rather than a risk averse approach,[242] it is understood that a human rights-based approach is not a "silver bullet" for addressing all possible harms that arise from the control of hate speech by intermediaries.[243] A margin of error must be assumed and a human rights-based approach must be concerned with open and transparent ways to manage the act of balancing the freedom of expression with the need to protect users from harm. Clear grievance, appeal and remedy mechanisms from content moderation decisions as guided by the UNGP and modelled on the DSA are therefore a necessary additional safeguard for the framework envisioned in this work.[244]

The proposed framework does not claim to provide solutions to all the human rights issues raised by the content moderation activities of intermediaries. It does, however, propose an approach for situating the regulation of hate speech by intermediaries within the international human rights law framework on the freedom of expression and addressing the issues identified in this work in a manner that gives due regard to the nuance of regulating online hate speech at a regional level. The application of international human rights norms and standards to the control of hate speech by intermediaries also diminishes the likelihood of content moderation being used to censor legitimate expression.[245]

E. CONCLUSION

The DSA presents a laudable attempt to address the concerns raised by the content moderation activities of intermediaries particularly as regards the control of hate speech and the protection of the freedom of expression. However, it fails to adopt a contemporary definition of hate speech and maintains the definition adopted in the 2016 Code of Conduct. Additionally, while setting guidelines

[240] N. Couldry and U. A. Mejias, "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject" (2019) 20 *Television and New Media* 336; C. Chanon et al, "From Extractivism to Global Extractivism: the evolution of an organizing concept" (2022) 49 *Journal of Peasant Studies* 760.

[241] UNGP, Principle 11.

[242] Machado and Aguiar, *supra* note 60 at 251.

[243] Sander, *supra* note 61 at 968.

[244] UNGP, principle 22, 29.

[245] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/38/35 (6 April 2018) at 15.

on the communication and enforcement of terms and conditions, it does not address the process by which such terms and conditions are developed by intermediaries. Furthermore, it fails to clearly state the human rights obligations of intermediaries opting for vague language which is open to interpretation. While setting impressive transparency and accountability safeguards, it leaves the job of determining the application of human rights concepts such as "proportionality" in online speech regulation to intermediaries.

There are substantive and procedural challenges inherent to regulating speech in a digital world such as definitional disparities, the complexity of online speech, the shortfalls inherent in algorithmic regulation and the issues arising from commercial content moderation, particularly in a regional context. To address some of these challenges, this paper proposes a non-binding human rights-based framework for the African region as a preliminary step towards regional consensus based on lessons from the EUs own progression towards adopting the DSA. It calls for protection of the freedom of expression by embedding international human rights law standards, *ubuntu,* and transparency and accountability into the regulation of hate speech online. Such a framework seeks to address the concerns not only in relations between the state and the individual as rights holder, but between the individual and intermediaries and between intermediaries and states. The proposed framework seeks to preserve the digital public sphere and ensure the promotion and protection of the freedom of expression in the regulation of hate speech online.