UNIVERSITY OF CAPE TOWN

SCHOOL OF ADVANCED LEGAL STUDIES

FACULTY OF LAW

Do global legal frameworks hold social media platforms accountable for hosting content that incites violence?

LLM Minor Dissertation

Master's Human Rights Law PBL 5626W

by

Ruvenna Samantha Rubenstein

RBNRUV001

Supervisor: ***Salona Lutchman***

29 November 2024 **Word Count: 24,822**

Research dissertation presented for the approval of the Senate in fulfillment of part of the requirements for the LLM in approved courses and a minor dissertation/ research paper. The other part of the requirement for this qualification was the completion of a programme of courses.

DECLARATION

I hereby declare that I have read and understood the regulations governing the submission of dissertations/ research papers, including those relating to length and plagiarism, as contained in the rules of this University, and that this dissertation/ research paper conforms to those regulations.

Signed by candidate

Ruvenna Samantha Rubenstein

Navigating Social Media Accountability: Do global legal frameworks hold social media platforms accountable for hosting content that incites violence?

Research abstract

The digital era has witnessed an unprecedented expansion in social media platforms' use, influence, and societal impact.[1] Sixty percent of the global population uses social media, with the daily exchange of messages reaching into the billions.[2] As of 2023, Facebook boasts 2.98 billion monthly active users,[3] YouTube exceeds 2.68 billion users,[4] and X (formerly Twitter) had 450 million users.[5] These platforms offer users unrestricted capacity for expressing views and communication, often with minimal (though not constant) oversight while facilitating the concealment of user identities.[6] While this technological advancement has opened new avenues for global connectivity and communication, it has also given rise to an alarming increase in the spread of hate speech.[7] In the last twenty years, these online platforms have evolved into environments where hateful narratives and stereotypes flourish unchecked, primarily aimed at marginalized groups, leading to increased communal violence, ethnic cleansing, and even genocide.[8] Major platforms such as Facebook, X, and YouTube have been criticized for failing to remove harmful content promptly and effectively, and for mistakenly removing content that does not breach their policies.[9]

This research endeavours to comprehensively investigate the accountability of social media platforms in addressing and mitigating the impact of hate speech that fuels acts of violence within the public sphere. Legal, ethical, and technological perspectives will be considered to

---

[1] Saurwein, F, and Spencer-Smith, C. "*Automated Trouble: The Role of Algorithmic Selection in Harms on Social Media Platforms*." (2021), *Media and communication (Lisboa)* Vol 9, Issue 4, pg. 222.
[2] Jikeli, G & Soemer K 'The value of manual annotation in assessing trends of hate speech on social media: was antisemitism on the rise during the tumultuous weeks of Elon Musk's Twitter takeover?' (2023) 6(2) *Journal of Computational Social* Science, pg. 945.
[3] Legal Resources Centre (LRC) A Critical Analysis of Content Moderation Policies and the Impact of Spreading Violence, Hatred & Disinformation in the Global South (2023) pg. 4 available at https://lrc.org.za/wp-content/uploads/LRC-CONTENT-MODERATION-RESEARCH-REPORT.pdf.
[4] *Ibid* at pg. 13.
[5] *Ibid* at pg. 25.
[6] Assimakopoulos S, Baide FH, Millar S (eds) *Online Hate Speech in the European Union, a Discourse-Analytic Perspective*, Springer Open, 2017, pg. 11, doi.org/10.1007/978-3-319-72604-5, available at https://library.oapen.org/bitstream/handle/20.500.12657/27755/1002250.pdf;jsessionid=8D114A5CB2B9462E038AF42E3ED7576B?sequence=1
[7] *Ibid.*
[8] Nourooz Pour H 'Transitional Justice and Online Social Platforms: Facebook and the Rohingya Genocide' (2023) 31(2) *International journal of law and information technology*, pg. 96.
[9] Hatano A, *'Regulating Online Hate Speech through the Prism of Human Rights Law: The Potential of Localised Content Moderation'* (2023) 41(1), *The Australian Year Book of International Law Online*, pg. 130.

examine the responsibilities borne by social media platforms in moderating user-generated content. A detailed analysis of existing legal frameworks, both national and international, governing hate speech and its consequences will be conducted to evaluate whether social media platforms are held accountable for content that incites violence. A comparative analysis of diverse social media platforms will be integral to this research, considering variations in policies, enforcement mechanisms, and responsiveness to instances of hate speech inciting violence. Case studies will be examined to illustrate specific incidents, shedding light on the challenges faced by social media platforms and the repercussions of inadequately addressing hate speech. This research aims to determine the legal responsibilities and accountability of social media platforms for hosting content that incites violence and examines whether the current measures are sufficient in addressing this critical issue.

Table of Contents

CHAPTER ONE
INTRODUCTION

1.1 Introduction

'So starting from now, we are the god of death for all (of them),'[10] 'Don't catch them alive.'[11] 'We must fight them the way Hitler did the Jews, damn kalars!'[12] 'Cut off those necks of the sons of the dog and kick them into the water.'[13] These are merely a few instances of the hate speech that spread on Facebook targeting the Rohingya, a minority group in Myanmar. The spread of hate speech on social media, namely Facebook, coincided with increased violence and human rights atrocities committed against the Rohingya.[14] Facebook was accused of disseminating intense hate speech that incited violence against this vulnerable minority group.[15] The United Nations' independent international fact-finding mission confirmed that Facebook had been an effective tool for the spread of hate in Myanmar.[16] This begs the question of whether social media platforms have a legal responsibility and accountability when content shared on their public platform causes and results in real-world violent events. With the prevailing toxicity surrounding the proliferation of hate speech on social media, this research aims to assess whether social media platforms are held accountable for hate speech that incites violence.

According to the Special Rapporteur report on minority issues for the period 2022-2023, hate speech against minorities in India for the period of 2014-2018 had increased by 786 percent.[17] The United States recorded its highest level of antisemitic incidents, and during this same period, islamophobia has also increased in many areas.[18] Hate speech often serves as a catalyst for the commission of hate crimes, creating a perilous

---

[10] Associated Press 'Myanmar: Facebook accused of letting hate speech thrive' CBS News 14 October 2023, available at https://www.cbsnews.com/news/myanmar-facebook-violent-hate-speech-thrives-ap-report/
[11] *Ibid.*
[12] Reuters 'Inside Facebook's Myanmar Crisis' Reuters 15 August 2018, available at https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/
[13] *Ibid.*
[14] Whitten-Woodring J et al 'Poison If You Don't Know How to Use It: Facebook, Democracy, and Human Rights in Myanmar' (2020) 25(3) *International Journal of Press/Politics* pg. 410.
[15] *Ibid.*
[16] United Nations Human Rights Council *Report of the Special Rapporteur on minority issues* (Fernand de Varennes) (3 March 2021) A/HRC/46/57, pg. 8 available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/054/14/PDF/G2105414.pdf.
[17] United Nations General Assembly *Minority Issues, Note by the Secretary-General* (16 August 2023) A/78/195 pg. 7 Available at https://documents-dds-ny.un.org/doc/UNDOC/GEN/N23/241/96/PDF/N2324196.pdf?OpenElement
[18] *Ibid* at pg. 8.

cycle wherein discriminatory rhetoric fosters and incites actual acts of violence or harm.[19] As highlighted in a submission to the Special Rapporteur, 'the Holocaust started not with the gas chambers, but with hate speech against a minority.'[20] There are instances in France where social media posts ultimately led to attacks on the Roma minority, and incidents where misinformation on social media caused several deaths against the Muslim minority in Sri Lanka.[21] An ongoing concern revolves around whether social media platforms should face consequences for providing a public platform that allows content that incites violence, genocide, and hostility.[22] Social media platforms that enable the dissemination of hateful and vile content act as potential catalysts for violence.[23] This stresses the need to thoroughly examine their responsibilities and accountability.

1.2 Historical background

While social media platforms offer tools for connecting with new people, developing relationships, sharing ideas, self-promotion, and self-expression, they have also created new channels for expressing hatred based on physical appearance, race, ethnicity, and gender.[24] Social media platforms have significantly amplified negative behaviour and hatred aimed towards individuals or groups by providing a public platform,[25] which can lead to serious consequences.[26]

On a societal level, hate speech has played a role in increasing tensions in communities, which has, in certain cases, led to violence.[27] The historical background of incidents where hate speech on social media played a role in real-world violence or contributed to the exacerbation of existing conflicts will be detailed and considered. A significant

[19] United Nations Human Rights Council *Report of the Special Rapporteur on minority issues* (Fernand de Varennes) (3 March 2021) A/HRC/46/57, pg. 7 available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/054/14/PDF/G2105414.pdf.
[20] *Ibid.*
[21] *Ibid at pg. 8.*
[22] *Ibid.*
[23] *Ibid.*
[24] Olteanu A, Castillo C, De Cristofaro E, & Varshney K, 'The Effect of Extremist Violence on Hateful Speech Online' *Proceedings of the Twelfth International Conference on Web and Social Media (ICWSM),* Stanford University California 2018, pg. 223.
[25] Keipi, T., Näsi, M., Oksanen, A., & Räsänen, P, "*Online Hate and Harmful Content: Cross-National Perspectives"* (1st ed.), Routledge, (2016) pg. 53.
[26] Olteanu A, Castillo C, De Cristofaro E, & Varshney K, *'*The Effect of Extremist Violence on Hateful Speech Online*' Proceedings of the Twelfth International Conference on Web and Social Media (ICWSM*), Stanford University California 2018 pg. 221.
[27] *Ibid.*

surge in negative articles about Jews in 1932 corresponds with a gross and substantial rise in antisemitic incidents in both Germany and Romania, as reported by multiple sources.[28] Joseph Goebbels, Hitler's propaganda minister, used modern media, such as films and radio, posters, and newspapers, to incite hatred.[29] This hate propaganda fuelled a climate of violence and intolerance that played a central role in enabling the extermination of the Jews in the Holocaust.[30] Goebbels further described radio as 'the most important instrument of mass influence that exists anywhere.'[31] Notably, in Rwanda in 1994, the radio was used to broadcast extreme and inciting messages against the Tutsi minority.[32] The radio station in Rwanda, being the dominant means of communication, broadcast messages calling for the extermination of the Tutsi minority and played a significant role in escalating violence against Tutsis.[33] Research indicates that 10% of the violence in the Rwanda genocide is attributable to the radio.[34] South Africa is another African country where content posted on social media platforms intensified real-world violence. In 2021, after the arrest of former President Jacob Zuma, his supporters used social media extensively[35] to show contempt for the social and economic disparity that instigated the violent protests.[36] In this instance, social media platforms were described as a 'dangerous tool'[37] used to incite violence and fuel the attacks.[38] Additionally, in Myanmar, as described above, inflammatory and vile Facebook posts inciting violence against the Muslim population played a significant role in fuelling violence against them.[39]

---

[28] Oberschall, A, 'Propaganda, hate speech and mass killings' in Oberschall A (ed) *Propaganda, War Crimes Trials and International Law* Chapter 5, (2012), pg. 176 available at http://francegenocidetutsi.org/OberschallPropagandaHateSpeechAndMassKillings2012.pdf.

[29] The Wiener Holocaust Library *The Holocaust Explained: The Nazi Rise to Power*, available at https://www.theholocaustexplained.org/the-nazi-rise-to-power/the-nazi-rise-to-power/propaganda/.

[30]Timmermann, KW 'The Relationship between Hate Propaganda and Incitement to Genocide: A New Trend in International Law Towards Criminalization of Hate Propaganda?' (2005) 18(2) *Leiden journal of international law* Pg 276.

[31] Yanagizawa-Drott, D, 'Propaganda and Conflict: Evidence from the Rwandan Genocide' (2014) 129(4) *The Quarterly journal of economics* pg. 1948.

[32] *Ibid.*

[33] *Ibid.*

[34] *Ibid.*

[35]The Presidency Republic of South Africa *Report of the Expert Panel into the July 2021 Civil Unrest* (29 November 2021) available at https://www.thepresidency.gov.za/sites/default/files/2022-05/Report%20of%20the%20Expert%20Panel%20into%20the%20July%202021%20Civil%20Unrest.pdf

[36] Karombo T 'South Africa goes after social media as it cracks down on looting and protests' *Quartz Africa* (14 July 2021) available at https://qz.com/africa/2033328/south-africa-to-monitor-social-media-as-protests-rock-the-country.

[37] *Ibid.*

[38] *Ibid.*

[39] Whitten-Woodring J et al 'Poison If You Don't Know How to Use It: Facebook, Democracy, and Human Rights in Myanmar' (2020) 25(3) *International Journal of Press/Politics* pg. 410.

These incidents, amongst others, prompted various lawsuits against social media platforms and triggered the commitment of social media platforms to regulate and aim to eliminate hateful content on their platforms. Taking into consideration the historical context and the ongoing issue of hate speech that incites violence on social media, should social media platforms be held accountable, and what responsibilities do they have in moderating and regulating content?

1.3 Hate speech that incites violence

To determine a social media platform's accountability, it is crucial to define what constitutes hate speech and, specifically, hate speech that results in the incitement of violence. In navigating the complex landscape of this online hate speech, it becomes apparent that there is no universally agreed-upon definition of hate speech.[40] In the 1980s, legal theorists coined the term "hate speech" in reference to expressions related to race.[41] Legal theorists often disagree with philosophers on how to define hate speech.[42] The legal definition of hate speech also differs across jurisdictions.[43] Facebook defines hate speech as a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability.'[44] According to the Council of Europe hate speech is defined as 'all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.'[45] Although the International Convention on the Elimination of Racial Discrimination (hereinafter "ICERD") does

[40] MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O 'Hate speech detection: Challenges and solutions' (2019) 14(8) *PLoS one* Pg 2.
[41] Lepoutre M et al 'What Is Hate Speech? The Case for a Corpus Approach' (2023) 18(2) *Criminal Law and Philosophy*, pg. 2.
[42] *Ibid.*
[43] Olteanu A, Castillo C, De Cristofaro E, & Varshney K, "The Effect of Extremist Violence on Hateful Speech Online "Proceedings of the Twelfth International Conference on Web and Social Media (ICWSM), 2018 pg. 2.
[44] MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O, "*Hate speech detection: Challenges and solutions*", (2019), Pg 2.
[45] Olteanu A, Castillo C, De Cristofaro E & Varshney K 'The Effect of Extremist Violence on Hateful Speech Online', paper presented at the *Proceedings of the Twelfth International Conference on Web and Social Media (ICWSM)* Stanford University California 25- 28 June 2018, pg. 2 available at https://ojs.aaai.org/index.php/ICWSM/issue/view/270

not explicitly define hate speech, Article 4 condemns the spreading of racial hatred propaganda and incitement to racial discrimination.[46] Furthermore, it mandates State parties to take legal measures against the dissemination of such ideas, incitement to racial discrimination, and acts of violence based on racial or ethnic grounds.[47]

The phrase "hate speech's" inherent weakness is that it is often dismissed as mere speech.[48] This sabotages effective responses to harm caused by speech that incites violence, discriminates against vulnerable groups, or suppresses marginalized voices.[49] Incitement is an inchoate crime, meaning that the occurrence of the incited action is not a prerequisite for the offense.[50] As a result, it is essential to establish a plausible connection between the spoken words and potential undesirable consequences. Various jurisdictions have differing views on the specific nature of this connection,[51] and establishing this connection may prove to be a complex task. The varying interpretations of hate speech that incites violence across jurisdictions can create challenges in enforcement, interpretation, and international cooperation, and this argument will be considered and explained.

1.4 Social Media Platforms' Accountability, Policies, Regulations, and Global Legal Frameworks

Holding social media platforms accountable for content on their platforms that incites violence involves a combination of legal, regulatory, and platform-specific measures. This research embarks on a thorough investigation into the legal frameworks governing hate speech that incites violence committed by users on social media platforms, and it delves into the interplay between social media platform policies and the broader spectrum of national and international regulations.

---

[46] International Convention on the Elimination of All Forms of Racial Discrimination, United Nations General Assembly resolution 2106 (XX) of 21 December 1965, entry into force, 4 January 1969, Article 4.
[47] *Ibid.*
[48] United Nations General Assembly *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (9 October 2019) A/74/486 available at https://documents-dds-ny.un.org/doc/UNDOC/GEN/N19/308/13/PDF/N1930813.pdf?OpenElement
[49] *Ibid.*
[50] Southern Africa Litigation Centre and Media Legal Defence Initiative *Freedom of Expression: Litigating Cases of Limitations to the Exercise of Freedom of Speech and Opinion Litigation* Manual Series (2016) Chapter 8, pg. 72, available at https://www.southernafricalitigationcentre.org/wp-content/uploads/2017/08/Chapter-8.pdf
[51] *Ibid.*

The regulation of hate speech is commonly framed by academics, civil society, and international organizations as a delicate equilibrium between the principles of free speech and other fundamental values, such as freedom from discrimination and human dignity.[52] To balance this delicate equilibrium, the content posted on social media platforms is subject to regulations through both internal policies and legal frameworks, which need to ensure that there is a balance between the preservation of freedom of expression and the prevention and prohibition of hate speech.[53]

Social media platforms are obligated to comply with legal regulations in the jurisdictions where they provide their services.[54] As Facebook, X, and YouTube are based in the United States, they are also subject to U.S. law.[55] Ensuring legal compliance can become complex where the applicable state laws are ambiguous, open to diverse interpretations, or inconsistent with human rights law.[56] According to submissions made to the Special Rapporteur in its 2021 report on minority issues, there is often a lack of enforcement of restrictions on hate speech on social media.[57] There are instances where there is no data on hate speech cases, existing legislation against hate crimes is not being utilized, or where it is deemed too challenging or unclear for successful prosecution.[58]

As governments and societies grapple with the imperative to regulate hate speech on social media platforms, an equally pressing concern emerges in ensuring that these platforms take decisive steps towards accountability, not only for hate speech but also

---

[52] Asogwa N & Ezeibe C 'The state, hate speech regulation and sustainable democracy in Africa: a study of Nigeria and Kenya' (2023) 20(3) *African Identities* pg. 25.
[53] O'Regan C 'Hate Speech Online: An (Intractable) Contemporary Challenge?' (2018) 71(1) *Current legal problems,* pg. 28.
[54] United Nations General Assembly, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Note by the Secretariat, (A/HRC/38/35)*. UN. 6 April 2018 para 22. Available at https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement.
[55] Gelashvili T *Hate Speech on Social Media: Implications of private regulation and governance gaps* (Master's Thesis, Faculty of Law, Lund University, (2018) available at https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8952399&fileOId=8952403
[56] United Nations General Assembly, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Note by the Secretariat, (A/HRC/38/35)*. UN. 6 April 2018 para 23. Available at https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement.
[57] United Nations Human Rights Council *Report of the Special Rapporteur on minority issues* (Fernand de Varennes) (3 March 2021) A/HRC/46/57., available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/054/14/PDF/G2105414.pdf.
[58] *Ibid.*

for crimes facilitated or enabled through their online spaces. Platform-enabled crimes refer to crimes not directly committed by social media companies that run social media platforms but are indirectly enabled or conducted by users through their online platforms.[59] Accountability for platform-enabled crimes differs based on the nature of the crime.[60] For purposes of this research, the platform-related crime examined refers to content that incites violence. Until recently, accountability discussions for platform-enabled crimes focused on two imperfect legal frameworks: international law or business and human rights.[61] International criminal law addresses individual liability for international crimes but lacks provisions for holding entities accountable.[62] Similarly, the law on state responsibility can impose civil liability on state entities but not companies.[63] Following incidents of violence stemming from social media posts, the primary emphasis is initially placed on the individuals responsible for posting the content, but complete accountability should include the social media platform that hosted the content.[64] It is crucial to assign accountability to social media platforms for their facilitative role in hosting content that incites real-world violence for two reasons. First, a social media post in isolation may not cause harm, but the cumulative impact of numerous social media posts can escalate and potentially incite violence.[65] Consequently, pursuing an individual for accountability would be meaningless. Second, holding social media platforms accountable may pave the way for essential changes in practices aimed at preventing harm arising from social media posts.[66]

1.5 Problem statement and research question of the minor dissertation

The research question of the minor dissertation is whether 'global legal frameworks hold social media platforms accountable for hosting content that incites violence?' Given that social media platforms operate globally across multiple jurisdictions with varying legislative systems,[67] this research will examine the legal frameworks that

[59] Hamilton R J 'Platform-enabled Crimes: Pluralizing Accountability when social media companies enable perpetrators to commit atrocities' (2022) 63(4) *Boston College Law Review,* pg. 1355.
[60] *Ibid* at pg. 1357.
[61] *Ibid* at pg. 1367.
[62] *Ibid* at pg. 1367.
[63] *Ibid* at pg. 1367.
[64] *Ibid* at pg. 1366.
[65] *Ibid* at pg. 1367.
[66] *Ibid.*
[67] Bayer J, Holznagel B, Korpisaari P & Woods L (eds) *Perspectives on Platform Regulation: Concepts and Models of Social Media Governance Across the Globe* vol 1 (2021) pg. 147 doi.org/10.5771/9783748929789

govern hate speech inciting violence, including international law, regional regulations, and domestic laws from selected jurisdictions. Additionally, the hate speech and violent content policies of key social media platforms, including their content moderation mechanisms, will be explored. Finally, illustrative case studies will be researched to assess the effectiveness of these legal frameworks and platform policies in addressing content that incites violence and holding social media platforms accountable.

1.6 Justification for the research

The potential link between online content and real-world violence necessitates the need to assess whether global legal frameworks effectively address these emerging challenges. It is essential to examine the effectiveness of existing legal frameworks in regulating social media platforms, as social media plays an increasingly significant role in shaping public discourse. This research aims to inform policymakers, legal professionals, and the public about the strengths and shortcomings of current legal frameworks and the policies and regulations that govern social media platforms. The research article seeks to contribute to discussions on refining regulations to ensure the public's well-being.

1.7 Research Methodology

The research methodology used to explore the accountability of global legal frameworks concerning social media platforms hosting content that incites public violence will primarily be desktop research. Academic literature, legal analysis, case law analysis, and international and regional laws will be used to provide a comprehensive understanding of the research.

1.8 Scope and limitations of the research

1.8.1 Social media platforms

The investigation into the accountability of social media platforms will be limited to three social media platforms, namely YouTube, Facebook, and X (previously known as Twitter), and will involve an examination of each of their content moderation regulations and policies with specific reference to hate speech that incites violence. In a 2019 joint statement, X, YouTube, and Facebook committed to upholding the Christchurch Call to Action to Eliminate Terrorist and Violent Extremist Content

Online (hereinafter the Christchurch Call).[68] The Christchurch Call community was created after a terrorist attack targeting the Muslim community in New Zealand was broadcast live on social media in 2019.[69] The Christchurch Call is a group consisting of Governments, civil society, and online service providers uniting in their commitment to combat online terrorist and violent extremist content.[70] However, despite Facebook, YouTube, and X's commitment to the Christchurch Call, most recent reports suggest that social media platforms are generally failing to act on hate speech targeting minorities.[71]

1.8.2 Legal Framework jurisdictions

The research will further narrow its focus to the following jurisdictions, namely, the United Kingdom, Germany, the United States of America, and South Africa, and their respective legal frameworks governing hate speech inciting violence will be assessed. These countries have been selected as they offer a diverse perspective. The United States is the birthplace of major social media platforms, and as a result, their data operations are predominantly conducted within the USA, subjecting them primarily to U.S. legal jurisdiction.[72] Facebook, based in Menlo Park, California, United States, is currently the largest social media platform in the world.[73] YouTube, with one billion visits a month, is based in San Bruno, California, United States, and Twitter is based in San Francisco, California, United States.[74] USA is further characterized by a notably expansive free speech tradition entrenched in the First Amendment,[75] leaving platforms

---

[68] Christchurch Call *Significant progress made on eliminating terrorist content online* (24 September 2019) available at https://www.christchurchcall.org/significant-progress-made-on-eliminating-terrorist-content-online/
[69] French Ministry for Europe and Foreign Affairs The Christchurch Call: What Progress Has Been Made? (12 May 2021) available at https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/the-christchurch-call-what-progress-has-been-made-12-may-2021
[70] Christchurch Call *The Christchurch Call Story* (15 May 2019) available at https://www.christchurchcall.org/the-christchurch-call-story/
[71] United Nations General Assembly *Minority Issues, Note by the Secretary-General* (16 August 2023) A/78/195.
[72] Gelashvili T *Hate Speech on Social Media: Implications of private regulation and governance* gaps (Master's Thesis, Faculty of Law, Lund University, 2018) pg. 58 available at https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8952399&fileOId=8952403
[73] World Atlas *Most Popular Social Media Networks in the World* Available at:https://www.worldatlas.com/articles/most-popular-social-media-networks-in-the-world.html
[74] *Ibid.*
[75] Chung M & Wihbey J 'Social Media Regulation, Third-Person Effect, and Public Views: A Comparative Study of the United States, the United Kingdom, South Korea, and Mexico' (2022) 26(8) *New media & society* pg. 4540.

largely unaccountable for harmful content.[76] In contrast, Germany implemented the Network Enforcement Act, commonly known as NetzDG, which requires social media platforms, including Facebook, X, and YouTube, to remove illegal content or face substantial fines.[77] This law applies to platforms with a substantial presence in Germany, including Facebook, X, and YouTube.[78] The United Kingdom has one of the highest percentages of social media users globally,[79] and on 26 October 2023, it assented to the Online Safety Act.[80] The Online Safety Act has been described as 'one of the most far-reaching attempts by a Western democracy to regulate online speech.'[81] The Online Safety Act applies to social media platforms that have a significant user presence in the UK.[82] Although social media platforms have initiated content moderation efforts globally, their level of activity seems relatively low on the African continent.[83] South Africa will serve as an illustrative case within the African continent, highlighting instances where appropriate content moderation seems to be lacking. This was evident when social media was used as a tool for inciting violence during the aftermath of former President Jacob Zuma's arrest.[84] With regard to the current South African legislation, the Prevention and Combating of Hate Crimes and Hate Speech Act was assented to by the President on 14 May 2024.[85]

1.9 Structure of the dissertation

The dissertation consists of five chapters. Chapter One is an introductory chapter; Chapter Two delves into the definition of hate speech; Chapter Three provides an

---

[76] Zurth P '*The German NetzDG as Role Model or Cautionary Tale? Implications for the Debate on Social Media Liability'* (2021) 31 *Fordham Intellectual Property Media & Entainment Law Journal* pg. 1092 Available at: https://ir.lawnet.fordham.edu/iplj/vol31/iss4/4

[77] *Ibid.*

[78] Kasakowskij T et al 'Network enforcement as denunciation endorsement? A critical study on legal enforcement in social media' (2020) 46 *Telematics and Informatics pg. 1.*

[79] The Global Statistics 'United Kingdom (UK) Social Media Statistics 2024 Most Popular Platforms' (2024) Available at https://www.theglobalstatistics.com/uk-social-media-usage-statistics/

[80] Online Safety Act 2023.

[81] 'Britain Passes Sweeping Online Safety Law' *New York Times* 19 September 2023, available at https://www.nytimes.com/2023/09/19/technology/britain-online-safety-law.html

[82] UK Department for Science, Innovation and Technology *Online Safety Act: Explainer* 92024) available at https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer

[83] Gore CD 'The politics of the internet and social media in Africa: three bases of knowledge for advancing research' (2023) 57(1) *Canadian Journal of African Studies / Revue canadienne des études africaines* pg. 212.

[84] Nxumalo L 'July unrest: Social media fuelled unrest' *IOL* 10 July 2022, available at https://www.iol.co.za/sunday-tribune/news/july-unrest-social-media-fuelled-unrest-481bbafe-1ce1-4ca7-87c9-7da6cfc652eb

[85] Republic of South Africa. Prevention and Combating of Hate Crimes and Hate Speech Act, Act No. 19 of 2019. Government Gazette No. 42794. Pretoria: Government Printer, 2019.

overview of the social media content moderation policies and regulations with specific reference to YouTube, Facebook, and X; Chapter Four analyses the global legal frameworks of the United Kingdom, United States of America, Germany and South Africa that govern social media platform accountability; Chapter Five provides a summary of the research findings together with potential recommendations for improvement and reform.

CHAPTER TWO
DEFINING HATE SPEECH

2.1 Introduction

This chapter explores the multifaceted nature of online hate speech that incites violence, dissects its various dimensions, and investigates a comprehensive framework for its definition. To provide an understanding of the diverse lenses through which hate speech is conceptualized, the chapter will research how online hate speech is defined from academic perspectives, international legal instruments, existing legal frameworks of the United Kingdom, United States of America, Germany, and South Africa, as well as outlining the hate speech policy definitions of YouTube, Facebook, and X.

2.2 Defining Hate Speech

The definitions of hate speech are often seen as unclear or inconsistent.[86] Assimakoupoulos et al. (2017) assert that hate speech lacks a 'universally accepted definition.'[87] Furthermore, legal interpretations showcase a range of conflicting depictions of the fundamental concepts and principles involved.[88] The diverse interpretations of hate speech across different contexts, cultures, legal systems, and social media policies highlight the challenges of defining hate speech and determining the appropriate responses.[89] As algorithms increasingly play a role in the initial analysis of content on social media and various online platforms, the need for a precise definition of hate speech becomes even more crucial to facilitate clear operationalization and ensure transparent procedures.[90] The research will begin with the exploration of the current definitions of hate speech within existing legal frameworks.

2.2.1 Legal Frameworks

To comprehensively understand the intricate nature of hate speech, it is imperative to explore and analyse its definitions within contemporary legal frameworks. This exploration encompasses international instruments that provide a broad understanding

[86] Hietanen M & Eddebo J 'Towards a Definition of Hate Speech—With a Focus on Online Contexts' (2023) 47(4) *Journal of Communication Inquiry pg. 441.*
[87] *Ibid.*
[88] *Ibid.*
[89] Hatano A 'Regulating Online Hate Speech through the Prism of Human Rights Law: The Potential of Localised Content Moderation' (2023) 41(1) *The Australian Year Book of International Law Online* pg. 155.
[90] *Ibid.*

of hate speech, as well as specific definitions adopted by various countries, reflecting the nuanced approaches each jurisdiction takes in addressing this complex issue.

2.2.1.1 International

While there is no single international instrument that universally defines hate speech, several international agreements and conventions address hate speech in various contexts, as detailed below.

2.2.1.1.1 The Universal Declaration of Human Rights[91]

The Universal Declaration of Human Rights (hereinafter "UDHR") was created by representatives with diverse legal and cultural backgrounds worldwide and is a significant document in the history of human rights.[92] 'Proclaimed by the United Nations General Assembly on December 10, 1948,'[93] the UDHR establishes a common standard for human rights, outlining fundamental freedoms to be protected for all people and nations.[94] Article 19[95] of the UDHR affirms everyone's right to freedom of expression[96] but does not explicitly prohibit hate speech.[97]

2.2.1.1.2 International Covenant on Civil and Political Rights [98]

The International Covenant on Civil and Political Rights (hereinafter "ICCPR") was adopted in 1966 and is currently ratified by 173 countries.[99] As opposed to the UDHR, the ICCPR is legally binding and contains a prohibition on hate speech.[100] According to Article 20,[101] of the ICCPR, 'Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility

[91] United Nations General Assembly. *Universal Declaration of Human Rights*, 10 December 1948.
[92] *Ibid.*
[93] *Ibid.*
[94] *Ibid.*
[95] Article 19, "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."
[96] *Ibid.*
[97] Mchangama J 'The Sordid Origin of Hate-Speech Laws' (2011) 170 *Policy Review* pg. 46.
[98] United Nations General Assembly. *International Covenant on Civil and Political Rights*. Adopted on 16 December 1966, entered into force 23 March 1976.
[99]Office of the United Nations High Commissioner for Human Rights 'Human Rights Indicators' available at: https://indicators.ohchr.org/.
[100] Mchangama J 'The Sordid Origin of Hate-Speech Laws' (2011) 170 *Policy Review* pg. 50.
[101] Article 20 1 "Any propaganda for war shall be prohibited by law. 2. Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."

or violence shall be prohibited by law.'[102] Although the ICCPR does not explicitly define hate speech, it emphasizes the restrictions on speech that encourage discrimination, hostility, or violence rooted in national, racial, or religious factors.

2.2.1.1.3 International Convention for the Elimination of All Racial Discrimination[103]

The International Convention for the Elimination of All Racial Discrimination (hereinafter "ICERD") was adopted by the United Nations in 1965.[104] ICERD expresses the commitment to adopt necessary measures to eliminate racial discrimination, prevent racist doctrines and practices, and promote a global community free from racial segregation and discrimination.[105] Article 4 of ICERD states:

"States Parties condemn all propaganda and all organizations which are based on ideas or theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form, and undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination and, to this end, with due regard to the principles embodied in the Universal Declaration of Human Rights and the rights expressly outlined in article 5 of this Convention."[106]

Article 4 of ICERD seeks to ensure that legal frameworks are in place to address and combat racial discrimination, both at the individual and organizational levels, and to create consequences for those who engage in or promote such discriminatory practices.[107] While ICERD doesn't explicitly define hate speech,

[102] United Nations General Assembly. *International Covenant on Civil and Political Rights*, Article 20. Adopted 16 December 1966, entered into force 23 March 1976.
[103] International Convention on the Elimination of All Forms of Racial Discrimination, adopted 21 December 1965, entered into force 4 January 1969.
[104] Mchangama J 'The Sordid Origin of Hate-Speech Laws' (2011) 170 *Policy Review* pg. 52.
[105] International Convention on the Elimination of All Forms of Racial Discrimination, adopted 21 December 1965, entered into force 4 January 1969, Preamble.
[106] International Convention on the Elimination of All Forms of Racial Discrimination, adopted 21 December 1965, entered into force 4 January 1969, Article 4.
[107] *Ibid.*

Article 4 addresses hate speech by focusing on the criminalization of expressions and actions that contribute to racial discrimination, hatred, and violence. The emphasis is on preventing and penalizing activities that could lead to or propagate racial discrimination rather than providing a comprehensive definition of hate speech itself.

2.2.1.1.3.1 General Recommendation No. 35 on the International Convention on the Elimination of All Forms of Racial Discrimination (hereinafter "General Recommendation No. 35")

Although the Convention does not explicitly employ the term 'hate speech,' General Recommendation No. 35 offers guidance on the definition of expressions that constitute racist hate speech within the Convention's framework.[108] As addressed in General Recommendation No.35, racist hate speech includes speech targeting 'indigenous peoples, descent-based groups, and immigrants or non-citizens, including migrant domestic workers, refugees, and asylum seekers, as well as speech directed against women members of these and other vulnerable groups.'[109] It also encompasses both direct and indirect forms of speech, as well as non-verbal expressions such as displaying racist symbols, regardless of whether they come from groups of individuals.[110]

2.2.1.1.4 European Convention on Human Rights[111]

Article 10[112] of the European Convention on Human Rights (hereinafter "ECHR") protects the right to freedom of expression. However, it does not

[108] UN Committee on the Elimination of Racial Discrimination (CERD), *General recommendation No. 35 : Combating racist hate speech*, CERD/C/GC/35, 26 September 2013.
[109] *Ibid.*
[110] *Ibid.*
[111] Council of Europe, *Convention for the Protection of Human Rights and Fundamental Freedoms*, as amended by Protocols Nos. 11 and 14, ETS 5, 4 November 1950.
[112] Article 10 "Freedom of expression 1. Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This Article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises. 2. The exercise of these freedoms, since it carries with it duties

define hate speech. Instead, it establishes the general principle of freedom of expression and allows for restrictions on this right in certain circumstances. These limitations on freedom of expression might pertain to hate speech and include *inter alia* the interests of national security, protection of public order, protection of morals and reputations of others, and public safety.[113]

2.2.1.1.5 Council of Europe's Additional Protocol to the Convention on Cybercrime concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems (hereinafter "Council of Europe's Additional Protocol to the Convention on Cybercrime ")[114]

The Council of Europe's Additional Protocol to the Convention on Cybercrime stressed the importance of ensuring a proper balance between freedom of expression and the effective prevention of racist and xenophobic acts.[115] Article 2[116] of the Council of Europe's Additional Protocol to the Convention on Cybercrime defines 'racist and xenophobic material'[117] as any written content, images, or expressions that advocate or incite hatred, discrimination, or violence against individuals or groups due to their race, colour, descent, national or ethnic origin, including situations where religion is used as a pretext for any of these discriminatory elements.[118]

---

and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary."

[113] Council of Europe, *Convention for the Protection of Human Rights and Fundamental Freedoms*, as amended by Protocols Nos. 11 and 14, ETS 5, 4 November 1950.

[114] Council of Europe, *Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems*, ETS 189, 28 January 2003.

[115] *Ibid.*

[116] Article 2 "Definition 1 For the purposes of this Protocol: "racist and xenophobic material" means any written material, any image or any other representation of ideas or theories, which advocates, promotes or incites hatred, discrimination or violence, against any individual or group of individuals, based on race, colour, descent or national or ethnic origin, as well as religion if used as a pretext for any of these factors. 2 The terms and expressions used in this Protocol shall be interpreted in the same manner as they are interpreted under the Convention."

[117] Council of Europe, *Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems*, ETS 189, 28 January 2003, Article 2.

[118] *Ibid.*

2.2.1.1.6 European Union's Code of Conduct on Countering Illegal Hate Speech Online (hereinafter "Code of Conduct")

In 2016, several major social media platforms, including Facebook, YouTube, and X, committed to the Code of Conduct regarding hate speech.[119] Their collective aim is to prevent the viral spread of hate speech.[120] This Code of Conduct defines hate speech as any action that publicly encourages violence or hostility towards a particular group of individuals or a member of such a group based on factors such as race, colour, religion, lineage, or national or ethnic background.[121]

2.2.1.1.7 UN Guiding Principles on Business and Human Rights

The global impact of multinational companies has led to the creation of the UN Guiding Principles on Business and Human Rights (hereinafter referred to as "UNGP").[122] While these principles are not legally binding, they establish standards that companies, including social media platforms, should follow.[123] This includes social media platforms respecting human rights.[124]

2.2.1.2 Regional

2.2.1.2.1 The African Charter on Human and Peoples Rights (hereinafter "The African Charter")[125]

The African Charter, also known as the Banjul Charter, does not explicitly mention hate speech by that specific term. However, the Charter contains provisions that relate to the principles underlying hate speech and its consequences. Article 9 of the African Charter guarantees the right to

---

[119] Hietanen M & Eddebo J 'Towards a Definition of Hate Speech—With a Focus on Online Contexts' (2023) 47(4) *Journal of Communication Inquiry*, Pg 444 available at https://journals.sagepub.com/doi/pdf/10.1177/01968599221124309.

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

[121]*Ibid.*

[122] United Nations High Commissioner for Human Rights *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (2011) available at https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

[123] *Ibid.*

[124]*Ibid.*

[125] Organization of African Unity (OAU), *African Charter on Human and Peoples' Rights* ("Banjul Charter"), 27 June 1981, entry into force 21 October 1986.

freedom of expression.[126] It states that 'every individual shall have the right to receive information'[127] and 'the right to express and disseminate his opinions within the law.'[128] Article 27 of the African Charter limits this right with a set of criteria designed to balance individual freedoms, such as the protection of national security, public order, and the rights of others.[129]

2.2.1.3 Specified Countries' Legal Frameworks

The definition of hate speech varies across legal systems, reflecting the diversity of cultural, social, and legal perspectives globally. Certain countries' criminal codes include definitions of hate speech in an effort to regulate harmful speech.[130] Legislation against hate speech exists in numerous European Union (hereinafter "EU") member states, though not uniformly across all.[131] Different countries criminalise different grounds for hate speech.[132] For instance, hate speech on the grounds of sexual orientation is criminalised in 20 EU member states, but only 6 EU member states criminalise hate speech on the grounds of age.[133] The definitions of hate speech, as detailed in the legislative frameworks of the UK, USA, Germany, and South Africa, will be detailed below.

2.2.1.3.1 United Kingdom

Since the enactment of the Race Relations Act[134] of 1965, the United Kingdom has had a distinct prohibition on hate speech.[135] The Race Relations Act forbade the incitement of discrimination or racial hatred.[136] Under their current law, the

---

[126] *Ibid, at* Article 9.
[127] *Ibid.*
[128] Organization of African Unity (OAU), African Charter on Human and Peoples' Rights ("Banjul Charter"), 27 June 1981, entry into force 21 October 1986, Article 9
[129] *Ibid*, at Article 27.
[130] Sellars, A, *Defining Hate Speech*, Berkman Klein Center Research Publication No. 2016-20, Boston Univ. School of Law, Public Law Research Paper No. 16-48, December 1, 2016, pg. 18 available at https://ssrn.com/abstract=2882244 or http://dx.doi.org/10.2139/ssrn.2882244.
[131] Hietanen M & Eddebo J 'Towards a Definition of Hate Speech—With a Focus on Online Contexts' (2023) 47(4) *Journal of Communication Inquiry* Pg 444 available at https://journals.sagepub.com/doi/pdf/10.1177/01968599221124309.
[132] *Ibid.*
[133] *Ibid.*
[134] Race relations Act,1965 c73.
[135] Sellars, A, *Defining Hate Speech*, Berkman Klein Center Research Publication No. 2016-20, Boston Univ. School of Law, Public Law Research Paper No. 16-48, December 1, 2016, pg. 19 available at https://ssrn.com/abstract=2882244 or http://dx.doi.org/10.2139/ssrn.2882244.
[136] *Ibid.*

Public Order Act[137] of 1986 (as amended), hate speech is described as 'threatening, or abusive words or behaviour, or displays any written material which is threatening, or abusive'[138] accompanied with the intention to stir up racial hatred.[139] Racial hatred is defined as 'hatred against a group of persons by reference to colour, race, nationality (including citizenship) or ethnic or national origins.'[140] Furthermore, any communications which are considered 'indecent and grossly offensive'[141] with the intention of causing distress or anxiety are criminalised in terms of Part 1 of the Malicious Communications Act.[142] Generally online hate speech is prosecuted under Section 127(1) of the Communications Act[143] 2003.[144] Section 127(1)[145] describes unlawful online content as electronic communication 'that is grossly offensive or of an indecent, obscene or menacing character.'[146] In *Connolly v DPP*[147], the term 'indecent or grossly offensive'[148] was examined when the defendant sent pictures of 21-week-old aborted foetuses to pharmacists who sold the morning-after pill.[149] The court deemed the photographs 'grossly offensive.'[150] The Online Safety Act, which aims to regulate the use of Internet services,[151] does not provide a singular definition of hate speech. Instead, it defines illegal content as 'content that amounts to a relevant offense.'[152]

---

[137] Public Order Act 1986, c. 64.
[138] *Ibid.*
[139] *Ibid.*
[140] *Ibid.*
[141] Malicious Communications Act 1988, c. 27.
[142] *Ibid.*
[143] Communications Act 2003, c. 21.
[144] Law Commission *Hate crime laws: Final Report* (2021) Law Com No 402 available at https://assets.publishing.service.gov.uk/media/61ba053ed3bf7f055eb9b8cf/Hate-crime-report-accessible.pdf
[145] Section 127 "Improper use of public electronic communications network
(1)A person is guilty of an offence if he—
(a)sends by means of a public electronic communications network a message or other matter that is grossly offensive or of an indecent, obscene or menacing character; or
(b)causes any such message or matter to be so sent."
[146] Communications Act 2003, c. 21, Section 127(1).
[147] *Connolly v DPP [2007] EWHC 237.*
[148] European Parliament, *Freedom of expression, a comparative-law perspective*, *The United Kingdom*, European Parliamentary Research Service, (October 2019) pg. 30.
[149] *Ibid.*
[150] *Ibid.*
[151] Online Safety Act 2023, c. 50 (UK).
[152] *Ibid.*

These categories form the framework for identifying and addressing hate speech in the UK. The legal definitions are designed to cover a spectrum of expressions that may incite discrimination, racial hatred, or cause distress based on the detailed categories.

2.2.1.3.2 United States of America

The United States, under the First Amendment of the Constitution, strongly protects freedom of speech.[153] As a result, hate speech is not a criminal offense unless it directly incites imminent violence or poses a true threat.[154] This means that a law governing hate speech must be narrowly tailored to serve a compelling state interest and must not be overbroad. The only form of hate speech that survives this scrutiny is incitement to violence, known as 'fighting words.'[155] To qualify as fighting words, the speech must directly incite or produce imminent lawless action and be likely to do so.[156] In general, the United States imposes restrictions on specific types of speech but does not have a universally agreed-upon definition of hate speech.

Brandenburg v Ohio[157] highlights the definition of hate speech that incites violence in the USA. According to this case law, the state cannot criminalize speech that advocates for violence or illegal actions unless it meets specific conditions.[158] For speech to be deemed criminally punishable, it must be directed towards inciting imminent unlawful conduct and have a high likelihood of actually causing such conduct.[159] This means that for speech to be considered beyond the protection of the First Amendment, it must be intended to incite immediate illegal actions, and there must be a clear risk that these actions will occur as a result of the speech. This framework ensures that only speech with a direct and imminent threat of incitement can be restricted.[160]

---

[153] University of Oxford, Oxford Pro Bono Publico *Comparative hate speech law: Annexure* Research paper prepared for the Legal Resources Centre, South Africa (March 2012) pg. 8 available at https://www.law.ox.ac.uk/sites/default/files/migrated/1a._comparative_hate_speech_annex.pdf.
[154] *Ibid.*
[155] *Ibid*.
[156] *Ibid.*
[157] *Brandenburg v. Ohio, 395 U.S. 444 (1969).*
[158] *Ibid.*
[159] *Ibid.*
[160] *Ibid.*

2.2.1.3.3 Germany

In Germany, hate speech is defined in the German Criminal Code (Strafgesetzbuch). Section 130(1)[161] of the German Criminal Code prohibits speech that incites hatred against national, racial, religious, or ethnic groups, as well as against specific segments of the population or individuals based on their membership in these groups.[162] It also prohibits speech that advocates violence or arbitrary actions against these groups or individuals or that violates their human dignity through insults, malicious statements, or defamation.[163] Further prohibitions include spreading portrayals of cruel, violent acts against human beings in a way that glorifies or diminishes their seriousness or violates human dignity and defaming beliefs, religious groups, or ideological organizations in a manner that jeopardizes public peace.[164]

2.2.1.3.4 South Africa

In South Africa, the Promotion of Equality and Prevention of Unfair Discrimination Act[165] (Equality Act) prohibits hate speech. Section 10 of the Equality Act states that it is unlawful for any individual to publish, spread, endorse, or communicate words related to 'race, gender, sex, pregnancy, marital status, ethnic or social origin, colour, sexual orientation, age, disability, religion, conscience, belief, culture, language, and birth'[166], or 'causes or perpetuates systemic disadvantage"[167], or "undermines human dignity'[168] if they show an intent to cause harm, be harmful, or encourage harm, as well as to promote or

---

[161] Section 130(1) "Whoever, in a manner suited to causing a disturbance of the public peace,
1. incites hatred against a national, racial, religious group or a group defined by their ethnic origin, against sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population, or calls for violent or arbitrary measures against them or
2. violates the human dignity of others by insulting, maliciously maligning or defaming one of the aforementioned groups, sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population
incurs a penalty of imprisonment for a term of between three months and five years."

[162] Criminal Code (StGB) Federal Law Gazette I (as last amended by Article 2 of the Act of 22 November 2021) Section 130(1) available at https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html.

[163] *Ibid.*

[164] O'Regan C 'Hate Speech Online: An (Intractable) Contemporary Challenge?' (2018) 71(1) *Current legal problems pg. 425.*

[165] Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000.

[166] Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000, Section 10.

[167] *Ibid.*

[168] *Ibid.*

spread hatred.[169] In *Qwelane v South African Human Rights Commission and Another*,[170] Qwelane likened gay and lesbian people to animals and claimed they were causing the decline of societal values in a 2008 Sunday Sun article.[171] The South African Human Rights Commission, confirmed by the Constitutional Court, found that this article incited harm, promoted hatred, and amounted to hate speech.[172] The Prevention and Combating of Hate Crimes and Hate Speech (Hate Speech Act)[173] defines hate speech as anyone who knowingly disseminates, promotes, advocates, provides, or communicates anything to one or more individuals in a way that could reasonably and intentionally cause harm or encourage harmful actions; and foster or spread hatred, based on one or more specified grounds.[174] In terms of the Hate Speech Act, the specified grounds include the following, 'albinism, ethnic or social origin, gender, HIV or AIDS status, nationality, migrant, refugee or asylum seeker status, race, religion, sex, sexual orientation, gender identity or expression or sex characteristics; or skin colour.'[175]

2.2.2 Academic Perspectives

Scholars approach the definition of hate speech from various perspectives, often influenced by their specific motivations.[176] Certain academics aim to make hate speech illegal and work to shape legislative and judicial responses through effective statutory language, while others do not advocate for legal consequences but rather to understand the concept of hate speech.[177]

In sifting through academic perspectives, a starting point is to examine Richard Delgado's persuasive article 'Words that Wound.'[178] In his article, Delgado emphasizes

---

[169] *Ibid.*
[170] *Qwelane v South African Human Rights Commission and Another CCT 13/20 [2021] ZACC 22.*
[171] *Ibid.*
[172] *Ibid.*
[173] Prevention and Combating of Hate Crimes and Hate Speech Act 15 of 2022, Government Gazette, Republic of South Africa.
[174] *Ibid.*
[175] Prevention and Combating of Hate Crimes and Hate Speech Act 15 of 2022, Government Gazette, Republic of South Africa, Definitions.
[176] Sellars A *Defining Hate Speech* Berkman Klein Center Research Publication No 2016-20, Boston Univ School of Law, Public Law Research Paper No 16-48 (1 December 2016) pg. 15 available at https://ssrn.com/abstract=2882244 or http://dx.doi.org/10.2139/ssrn.2882244.
[177] *Ibid.*
[178] Delgado R 'Words that wound A tort action for racial insults, epithets and name calling' (1982) 17 *Harvard Civil Rights-Civil Liberties Law Review.*

racism, advocating for legal repercussions against racist speech.[179] Delgado's definition of hate speech requires the plaintiff to prove the following: (1) the defendant's language was meant to insult the plaintiff based on race, (2) the plaintiff perceived it as racially derogatory, and (3) a reasonable person would view it as a racial insult.[180] Delgado's definition, when dissected, notably lacks specific content criteria but centers instead on intent, impact, and objective perception. More specifically, it scrutinizes the speaker's intention to 'demean through reference to race.'[181]

Mari J. Matsuda expands on Delgado's ideas, examining hate speech from a legal perspective.[182] Calvin Massey, on the other hand, states that hate speech encompasses any speech causing the harm typically associated with hate speech suppression: loss of self-esteem, economic and social marginalization, physical and mental strain, victim silencing, and effective exclusion from political participation.[183] Another scholar, Kenneth Ward, described hate speech as any expression with the intention to vilify, humiliate, or incite hatred.[184]

In 2014, Alice Marwick and Ross Miller examined multiple definitions extensively, including those proposed by Massey and Ward above.[185] They pinpointed three overarching aspects frequently employed to define hate speech: (1) content-focused, (2) intent-driven, and (3) harm-oriented elements.[186]

When scrutinizing definitions found in academic literature, the language utilized often carries a significant emotional charge and reflects a particular moral stance.[187] The risk of using emotionally charged and vague definitions, is that they could lead to politically sensitive or provocative opposition being labelled as hate speech, silencing dissenting

---

[179] *Ibid.*
[180] *Ibid at* pg. 179.
[181] Sellars A *Defining Hate Speech* Berkman Klein Center Research Publication No 2016-20, Boston Univ School of Law, Public Law Research Paper No 16-48 (1 December 2016) pg. 16 available at https://ssrn.com/abstract=2882244 or http://dx.doi.org/10.2139/ssrn.2882244.
[182] *Ibid.*
[183] *Ibid.*
[184] *Ibid at pg. 17.*
[185] *Ibid at pg. 18.*
[186] *Ibid.*
[187] Hietanen M & Eddebo J 'Towards a Definition of Hate Speech—With a Focus on Online Contexts' (2023) 47(4) *Journal of Communication Inquiry pg. 444.*

views.[188] The intricacy of what needs to be encompassed within the definition has led to either abstract or impractical definitions, and it seems, according to Papcunová et al. (2021), that the likelihood of determining a 'universal theoretical definition of hate speech'[189] is slim.[190]

Often, a broad definition is provided without a critical discussion, such as in the case of Nockleby, who provides a commonly used definition stating that hate speech is 'usually thought to include communications of animosity or disparagement of an individual or a group on account of a group characteristic such as race, colour, national origin, sex, disability, religion or sexual orientation.'[191]

Several academics like Sellars do not explicitly offer a definition, but based on a review of literature, imply 'some common themes'[192] for defining hate speech: (1) targeting a group or an individual as part of a group, (2) expressing hate through the content, (3) speech leading to harm, (4) intention to cause harm, (5) prompts harmful actions, (6) the speech is made public or directed at a group member, (7) the context allows for potential violent reactions, and (8) the speech lacks any constructive purpose.[193]

Others like Schmidt and Wiegand (2017) advocate for a comprehensive interpretation of hate speech, defining it as a wide-ranging term covering various forms of offensive content created by users.[194] Certain academics contend that hate speech can be deemed criminal in specific situations, even if the language used does not directly incite violence.[195] Conversely, others recognize that hate speech, which stops short of urging

[188] *Ibid.*
[189] *Ibid* at *pg. 445.*
[190] *Ibid.*
[191] *Ibid.*
[192] Sellars A *Defining Hate Speech* Berkman Klein Center Research Publication No 2016-20, Boston Univ School of Law, Public Law Research Paper No 16-48 (1 December 2016) pg. 25-30 available at https://ssrn.com/abstract=2882244 or http://dx.doi.org/10.2139/ssrn.2882244.
[193] Sellars A *Defining Hate Speech* Berkman Klein Center Research Publication No 2016-20, Boston Univ School of Law, Public Law Research Paper No 16-48 (1 December 2016) pg. 25-30 available at https://ssrn.com/abstract=2882244 or http://dx.doi.org/10.2139/ssrn.2882244.
[194] Vilar-Lluch S 'Understanding and appraising "hate speech"' (2023) 11(2) *Journal of Language Aggression and Conflict* pg. 2 available at :https://www.jbe-platform.com/content/journals/10.1075/jlac.00082.vil.
[195] Fino A 'Defining Hate Speech: A Seemingly Elusive Task' (2020) 18(1) *Journal of International Criminal Justice* pg. 33 available at https://doi.org/10.1093/jicj/mqaa023.

violence, may not be subject to criminalization. They suggest that nations should consider implementing a unified liability treaty regarding 'atrocity speech offenses.'[196]

It is clear from the above that academic perspectives on the definition of hate speech emphasize the complexity and challenges inherent in identifying and addressing this concept. While there is a consensus amongst the above academics that hate speech involves discriminatory language or behavior targeting individuals or groups based on certain protected characteristics or attributes, the precise delineation of what constitutes hate speech remains subject to debate and interpretation.

2.3. Conclusion

In conclusion, the examination of the definition of online hate speech in this chapter sheds light on its intricate and multifaceted nature. Delving into numerous international legal instruments shows a lack of a universal definition of hate speech. The UDHR fails to define hate speech. While some instruments, such as the ICCPR and ICERD, provide certain restrictions on speech that encourage discrimination or violence based on national, racial, or religious factors, they do not explicitly define hate speech. They focus rather on the consequences of such speech and aim to prevent its consequential effects. On the other hand, the Convention on Cybercrime and the Code of Conduct offer more specific definitions of hate speech, particularly in the context of online communication, but with limited discriminatory categories.

The analysis of hate speech legal definitions across different jurisdictions reveals notable discrepancies and nuances. In the United Kingdom, the emphasis is to prohibit speech that incites racial hatred, whereas the United States only imposes restrictions primarily when speech directly leads to violence. Germany's legal framework focuses on speech directed against specific groups based on national, racial, religious, or ethnic affiliations. In contrast, South Africa adopts a comprehensive approach, with a detailed list of prohibited grounds for hate speech. The consequence of these differing interpretations of hate speech across jurisdictions leads to inconsistencies in addressing and combating this harmful phenomenon. When legal definitions and approaches vary significantly from one jurisdiction to another, it leads to confusion, ambiguity, and challenges in enforcing laws and policies related to hate speech.

---

[196] *Ibid.*

This research has uncovered the complexity inherent in defining what constitutes hate speech. It appears that there is consensus within legal frameworks and academic circles that hate speech encompasses speech that discriminates against individuals or groups based on various categories. However, the lack of a uniform and clear definition poses significant challenges for policymakers, law enforcement agencies, and online platforms tasked with addressing and mitigating the spread of hate speech.

CHAPTER 3

Social Media Platforms' Policies on Hate speech that incites violence.

3.1 Introduction

Given the volume of content being shared, it is essential to have precise definitions of hate speech and violent or harmful material to ensure fair and consistent moderation.[197] This is illustrated in *Themel v Facebook*,[198] a matter in the Higher Regional Court in Munich, which ruled that Facebook unlawfully deleted a user's comment.[199] Although the platform categorized the comment as hate speech under its community guidelines, the court ruled it did not meet the legal threshold for hate speech.[200] This case underscores the challenges of an inconsistent definition and interpretation of hate speech within social media platforms' policies and legal frameworks.

Despite these inconsistencies, the daily decisions and determination on what content is acceptable and how to manage hate speech are ultimately determined by the social media platform, guided by their private regulations, such as terms of service and community guidelines.[201] As a result, the boundaries of freedom of expression and the right to equality are determined by these private entities, placing significant responsibility for upholding fundamental rights in their hands.[202]

Content moderation refers to the process by which platforms review content created by users on their platform and decide whether to leave it online or remove it.[203] The challenge with this process is enforcing a universal set of rules for speech without a universal definition of what speech should be permitted and managing this on a large scale.[204] As platforms wrestle with this challenge, several issues are becoming increasingly apparent. Despite their active efforts

---

[197] De Cook J R, Cotter K, Kanthawala S & Foyle K 'Safe from "Harm": The Governance of Violence by Platforms' (2022)14(1) *Policy & Internet* pg. 67.
[198] *Heike Themel v. Facebook Ireland Inc. 24.08.2018 - 8 W 1294/18*
[199] Columbia University Global Freedom of Expression, *Heike Themel v. Facebook Ireland Inc.*, available at https://globalfreedomofexpression.columbia.edu/cases/heike-themel-v-facebook-ireland-inc/
[200] *Ibid.*
[201] Gelashvili T *Hate Speech on Social Media: Implications of private regulation and governance gaps* (Master's Thesis, Faculty of Law, Lund University, 2018) pg. 52 available at https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8952399&fileOId=8952403
[202] *Ibid.*
[203] Kloncik, K. The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression (2020) 129(8) *The Yale Law Journal*, pg. 2427.
[204] *Ibid.*

to regulate content, academic research on content moderation has shown that platforms' policies often disproportionately affect marginalized communities that already experience stigma in society.[205] Furthermore, platforms' increased reliance on automated systems to moderate large volumes of content[206] raises growing concerns regarding implementation and fairness.[207] Automated hate speech moderation poses several challenges. One major issue is that users often find it difficult to hold these automated systems accountable or effectively appeal their decisions.[208] Another criticism is the system's limited inability to accurately process and understand all languages, particularly when it comes to capturing nuanced cultural or contextual meanings.[209] This can result in misinterpretations, with some harmful content remaining on the platform or non-violating content being incorrectly flagged.

Understanding social media platforms' policies is crucial to evaluating the consistency and fairness of their enforcement practices, particularly in terms of accountability. This chapter will focus on the hate speech and violent content policies of Facebook, YouTube, and X, including their definitions of hate speech, enforcement methods, and the extent to which they comply with legal frameworks.

3.2 Facebook

Facebook is a social networking site and is a public company.[210] The Facebook company was renamed Meta, which owns various social media platforms, including the Facebook platform,[211] referred to as Facebook throughout this research for ease of reference. From 2004 to 2010, Facebook lacked a formal content-moderation policy, with a small team following a vague set of guidelines to remove harmful content.[212] It wasn't until 2008 that the platform developed its first public 'Community Standards'[213] and a detailed internal 'Abuse

[205] Are C '"Dysfunctional" Appeals and Failures of Algorithmic Justice in Instagram and TikTok Content Moderation' (2024) *Information, communication & society* pg. 2.
[206] *Ibid* at pg. 3.
[207] *Ibid.*
[208] *Ibid.*
[209] Hatano A 'Regulating Online Hate Speech through the Prism of Human Rights Law: The Potential of Localised Content Moderation' (2023) 41(1) *The Australian Year Book of International Law Online pg.* 145.
[210] Underwood M 'Is Facebook a Private Company?' *Market Realist* 10 February 2023 available at https://marketrealist.com/p/is-facebook-a-private-company/.
[211] *Ibid.*
[212] Kloncik, K. The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression (2020) 129(8) *The Yale Law Journal*, pg. 2435.
[213] Kloncik, K. The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression (2020) 129(8) *The Yale Law Journal*, pg. 2435.

Standards'[214] document outlining enforcement.[215] As public pressure grew and European nations like Germany enforced stricter speech laws, Facebook's Community Standards became more aligned with European norms, yet largely lacked input from the Global South.[216] These global community standards specify the content and behaviour not allowed on Facebook's platform.[217] These standards serve as a guideline for the type of user-generated content that is permitted.[218] In developing these community standards and enforcing their implementation, Facebook consults with international human rights experts and other stakeholders.[219] Facebook has in place extensive technologies, systems, and controls to find and remove content that violates its policies, sometimes before it is disseminated.[220]

3.2.1 Hate Speech Policy

In terms of Facebook's hate speech policy, hate speech is strictly forbidden on its platform.[221] However, they acknowledge that people may share such content to condemn hate speech or bring attention to it.[222] Sometimes, slurs or similar language may be used self-referentially or in a way that empowers certain groups.[223] Facebook's policies attempt to accommodate these uses of language but require a clear indication of the user's intent.[224] If the intention behind the use of such language is unclear, Facebook may take action to remove the content.[225] In some cases, content that violates Facebook's community standards may be permitted if it is satirical and the offending elements are used to mock or critique something else.[226]

---

[214] *Ibid.*
[215] *Ibid.*
[216] Ibid at pg. 2437.
[217] Facebook *Understanding Community Standards* available at https://www.facebook.com/community/using-key-groups-tools/understanding-community-standards/
[218] Facebook *Facebook's Corporate Human Rights Policy*, March 2021, available at https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf.
[219] *Ibid.*
[220] Meta *How technology detects violations* available at https://transparency.meta.com/en-gb/enforcement/detecting-violations/technology-detects-violations/
[221] Meta *Facebook Community Standards: Hate Speech*, Meta Transparency Center, available at https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/
[222] *Ibid.*
[223] *Ibid.*
[224] *Ibid.*
[225] *Ibid.*
[226] *Facebook, Hate speech policy,* available at https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/.

Facebook's current policy on hate speech defines hate speech as direct attacks on individuals based on 'protected characteristics'[227], which include 'race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease.'[228] Direct attacks are defined as violent or dehumanizing language, detrimental stereotypes, declarations of inferiority, displays of contempt, disgust, or dismissal, the use of profanity, and demands for exclusion or segregation.[229] Facebook further disallows damaging and dangerous stereotypes described as dehumanizing comparisons historically utilized to target, intimidate, or marginalize specific groups, often correlated with offline acts of violence.[230]

Facebook divides hate speech content into three levels of harm spanning from the most severe to comparatively less.[231] The first tier encompasses the most harmful forms of hate speech, such as violent or dehumanizing speech targeting individuals or groups based on protected characteristics or immigration status.[232] Posts that are prohibited include advocating or supporting violence either through written statements or visual content as well as dehumanizing language or imagery that compares individuals or groups to insects, culturally stereotyped animals, or uses derogatory terms related to filth or disease.[233] Furthermore, it forbids portraying people as subhuman or predatory based on their characteristics, criminalizing entire groups, denying the existence of protected characteristics, perpetuating harmful stereotypes linked to violence or exclusion (such as Blackface or Holocaust denial), and mocking victims or events of hate crimes.[234]

The second tier involves less severe forms of hate speech, including generalizations and expressions of inferiority based on physical, mental, or moral characteristics.[235] It also includes derogatory expressions, expressions of contempt or hate, expressions of dismissal or disgust,

---

[227] *Ibid.*
[228] *Ibid.*
[229] *Ibid.*
[230] *Ibid.*
[231] Hietanen M & Eddebo J 'Towards a Definition of Hate Speech—With a Focus on Online Contexts' (2023) 47(4) *Journal of Communication Inquiry*, pg. 448 available at https://journals.sagepub.com/doi/pdf/10.1177/01968599221124309.
[232] Facebook, *Hate speech policy*, available at https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/.
[233] *Ibid.*
[234] *Ibid.*
[235] *Ibid.*

and certain types of cursing.[236] This includes derogatory terms like 'dumb,' 'stupid,' 'retarded,' 'crazy,' or 'insane,'[237]'coward,' 'liar,' or 'arrogant,' as well as expressions that demean or belittle others as 'freaks,' 'useless,' or 'worthless.'[238]

The third tier addresses hate speech involving segregation or exclusion based on protected characteristics.[239] This includes calls for action, intent statements, or advocacy supporting segregation or exclusion in political, economic, or social contexts.[240] Facebook has another category that requires further information gathering before a decision can be made.[241]

3.2.2 Violence and Incitement Policy

Facebook's violence and incitement policy aims to prevent potential offline harm related to Facebook content. This policy complements Facebook's Hate Speech Policy, which targets content that discriminates against or incites hatred toward individuals or groups based on protected characteristics. While the Hate Speech Policy focuses on preventing harmful speech based on identity and beliefs, the Violence and Incitement Policy specifically addresses content that incites or facilitates violence, including threats and calls for physical harm.[242] Therefore, while Facebook accepts that users might occasionally make non-serious threats when expressing disagreement, they take action to remove content that incites or promotes serious violence.[243] This involves removing harmful content, disabling accounts, and working with law enforcement when there is a real threat to public safety.[244] Facebook evaluates both the language and context to distinguish between casual statements and credible threats and considers additional factors such as the person's public visibility and associated safety risks when determining the credibility of a threat.[245]

3.2.3 Enforcement

3.2.3.1 Content Review

---

[236] *Ibid.*
[237] *Ibid.*
[238] *Ibid.*
[239] *Ibid.*
[240] *Ibid.*
[241] *Ibid.*
[242] Meta *Community Standards: Violence and Incitement* Meta Transparency Center, available at https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/
[243] *Ibid.*
[244] *Ibid.*
[245] *Ibid.*

Facebook's content moderation process involves a combination of technology and human review teams to manage millions of pieces of content daily.[246] Facebook utilizes artificial intelligence and machine learning tools to automatically detect and remove a significant portion of content that violates community standards, often before it is seen by users.[247] When technology alone cannot determine whether content breaches policies, it is reviewed by thousands of global moderators (with language expertise and knowledge of the subject matter) who assess each post individually.[248] Reviewers prioritize content based on severity, virality, and likelihood of policy violations.[249] In cases where context is critical, reviewers may receive additional information to help make informed decisions.[250] Despite Facebook confirming that moderators with sufficient language expertise are reviewing content, with a total of 220 countries having access to Facebook,[251] detection of hate speech content in many Indigenous languages proves extremely challenging and accordingly increases vulnerability to online hate.[252] In June 2023, the South African Legal Resource Centre and Global Witness investigated the effectiveness of various social media platforms' content moderation and review processes.[253] The investigation created fake accounts to post hate speech ads that violated Facebook's hate speech policy in English, Afrikaans, isiZulu, and isiXhosa.[254] Of the 38 ads, Facebook approved all except one ad in English and Afrikaans, which was still approved in isiZulu and isiXhosa.[255] This highlights the potential prejudice and inequities that can arise in multilingual societies when social media platforms fail to provide consistent and fair content moderation across all languages. Bias, including language bias, can affect content assessment regardless of whether the evaluation is conducted by a human or a machine.[256] This

---

[246] Meta *How Review Teams Work* Meta Transparency Center, available at https://transparency.meta.com/en-gb/enforcement/detecting-violations/how-review-teams-work/
[247] *Ibid.*
[248] *Ibid.*
[249] Meta *Prioritizing Content Review* Meta Transparency Center, available athttps://transparency.meta.com/en-gb/policies/improving/prioritizing-content-review/
[250] Meta *How Review Teams Work* Meta Transparency Center, available at https://transparency.meta.com/en-gb/enforcement/detecting-violations/how-review-teams-work/
[251] World Population Review *Facebook Users by Country*, available at https://worldpopulationreview.com/country-rankings/facebook-users-by-country
[252] Legal Resources Centre (LRC) *A Critical Analysis of Content Moderation Policies and the Impact of Spreading Violence, Hatred & Disinformation in the Global South* (2023) pg. 6 available at https://lrc.org.za/wp-content/uploads/LRC-CONTENT-MODERATION-RESEARCH-REPORT.pdf.
[253] *Ibid at pg. 40.*
[254] *Ibid.*
[255] *Ibid.*
[256] Diaz A & Hecht-Felella L, *Double Standards in Social Media Content Moderation*, Brennan Center for Justice, New York University School of Law (2021) pg. 11 available at https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf

discretion in determining whether content violates community standards can lead to both excessive removal of acceptable content and insufficient action against harmful content.[257]

3.2.3.2 Content Removal

In terms of Facebook's hate speech policy, any content found to violate Facebook's Community Standards, as listed above, will be removed.[258] The platform uses a strike system to track and manage content violations.[259] The number of strikes and the nature of the policy breached, along with a user's violation history, can lead to account restrictions or even disabling.[260] Marginalized communities are at greater risk for adverse content removal due to the higher error rates in algorithmic versus manual takedowns.[261] In this context, a significant instance took place in Nigeria in 2020, when Facebook's automated tools mistakenly removed posts containing the #EndSARS hashtag, which was being used to raise awareness about police brutality and violence in Nigeria.[262] The posts were incorrectly classified as misinformation related to COVID-19 and removed erroneously.[263] This emphasizes the importance of understanding local context and nuances to mitigate the risk of errors caused by automated processes for content removal.

3.2.3.3 Oversight Board and Appeal

In 2018, Facebook CEO Mark Zuckerberg proposed a new approach to ensuring accountability and legitimacy in content control and moderation on platforms like Facebook.[264] This led to the creation of the Oversight Board (hereinafter "the board"), established to provide an independent review of Facebook's most challenging content decisions.[265] The board represents the first significant attempt at independent adjudication of online speech at a large scale.[266] The core principle behind the creation of the board was that Facebook should not be the sole

---

[257] *Ibid.*
[258] Meta *Taking Down Violating Content* Meta Transparency Center available at https://transparency.meta.com/en-gb/enforcement/taking-action/taking-down-violating-content/
[259] *Ibid.*
[260] *Ibid.*
[261] Diaz A & Hecht-Felella L, *Double Standards in Social Media Content Moderation*, Brennan Center for Justice, New York University School of Law (2021) pg. 18 available at https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf
[262] *Ibid.*
[263] *Ibid.*
[264] Meta *Creation of the Oversight Board* Meta Transparency Center available at https://transparency.meta.com/en-gb/oversight/creation-of-oversight-board/
[265] *Ibid.*
[266] Klonick, K 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' *(2020) 129(8) The Yale law journal* at pg. 2499.

authority on crucial decisions regarding free expression.[267] The board commenced in May 2020, consisting of a diverse group of experts, including professors, journalists, and former heads of state,[268] and began reviewing cases in October 2020.[269] Users who disagree with Facebook's content decisions can appeal to the board using a reference number obtained through Facebook's appeals process.[270] The board reviews cases at its sole discretion and can issue binding decisions and recommendations on content policies and enforcement practices.[271] Initially, the board frequently overturned decisions, suggesting that Facebook's decision-making process is flawed and highlighting the need for appeals to rectify errors.[272] Despite this improvement to their appeals process, the accessibility to the appeals process to users from marginalised communities remains uncertain.[273]

A recent appeal case heard by the board was Facebook's failure to remove a video posted in Nigeria in December 2023.[274] The video showed two men allegedly being assaulted for being homosexual.[275] The video, spoken in Igbo, remained on Facebook for five months and gained 3.6 million views.[276] Facebook confirmed that they had misidentified the language spoken in the video as Swahili.[277] In determining whether the video should be removed, the board assessed both Facebook's policies and international legal frameworks.[278] The board confirmed the video violated Facebook's hate speech policy, violence, and incitement policy and was not consistent with Facebook's international human rights law responsibilities.[279] Despite 112 reports and three moderator reviews, the post was only removed after the Board's intervention.[280] This case highlights flaws in content moderation, cultural and linguistic bias,

---

[267] Meta *Creation of the Oversight Board* Meta Transparency Center available at https://transparency.meta.com/en-gb/oversight/creation-of-oversight-board/
[268] *Ibid.*
[269] *Ibid.*
[270] *Ibid.*
[271] Meta *Oversight Board Recommendation* Meta Transparency Center available at https://transparency.meta.com/en-gb/oversight/oversight-board-recommendations
[272]Diaz A & Hecht-Felella L *Double Standards in Social Media Content Moderation* Brennan Center for Justice, New York University School of Law (2021) pg. 18 available at https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf
[273] *Ibid.*
[274] Oversight Board *Homophobic Violence in West Africa, Full Case Decision* (2024) available at https://www.oversightboard.com/decision/fb-ouuwkhko/
[275] *Ibid.*
[276] *Ibid.*
[277] *Ibid.*
[278] *Ibid.*
[279] *Ibid.*
[280] *Ibid.*

risks to vulnerable communities, and the importance of external oversight and alignment with international human rights law.

3.2.3.4 Compliance with Legal Frameworks

If content on Facebook is reported as violating local laws but not Facebook's Community Standards, Facebook might limit access to that content only in the country where it's deemed illegal.[281]

Facebook's enforcement of its hate speech and violence and incitement policy involves rigorous review processes, including the new independent oversight board referred to above. Between April and June 2024, the platform acted on 7.2 million pieces of hate speech content.[282] 1.2 million of this actioned content was appealed by users, resulting in the restoration of 157,000 pieces of content after review.[283] During the same period, Facebook acted on 7.4 million pieces of violent and incitement content.[284] Of this actioned content, 1.3 million were appealed by users, resulting in the restoration of 225,000 pieces of content after review.[285] These figures provided by Facebook indicate extensive enforcement; however, internal documents leaked in November 2021 highlighted significant flaws in Facebook's moderation software, revealing a bias against minority voices.[286] Approximately 90 percent of the 'hateful' content removed criticized white individuals and/or men.[287] Further studies indicate that Facebook's content moderation disproportionately impacts marginalized groups.[288] While the platform has revised its race-blind moderation policies, which previously failed to account for systemic inequalities, concerns about ongoing discrimination remain.[289]

3.3 YouTube

Like Facebook, YouTube enforces community guidelines to manage the enormous volume of content uploaded to its platform. YouTube is an online video hosting platform founded in 2005,

---

[281] *Ibid.*
[282] Meta *Community Standards Enforcement Report: Hate Speech on Facebook* Meta Transparency Center available at https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook/
[283] *Ibid.*
[284] *Ibid.*
[285] *Ibid.*
[286] Griffin, R 'Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality' (2023) 2(1) *European Law Open* pg. 30.
[287] *Ibid.*
[288] *Ibid.*
[289] *Ibid.*

owned by Google, and is a private company.[290] With 500 hours of videos being uploaded every minute, YouTube must navigate a vast library of content.[291] To ensure compliance with community guidelines, YouTube utilizes a combination of advanced machine learning systems and community reporting.[292] Content flagged as potentially problematic is reviewed by expert teams, who then remove any material that violates these guidelines.[293] When using YouTube, users become part of a global community and are encouraged to report any content they believe violates these community guidelines.[294] Occasionally, content that might otherwise breach community guidelines may remain on YouTube as an exception if it has educational, documentary, scientific, or artistic (EDSA) value.[295] YouTube's hate speech and violent and graphic content policies will be detailed below.

3.3.1 Hate speech policy

YouTube defines hate speech as content that promotes violence or hostility towards individuals or groups based on certain attributes considered protected under its policy, such as 'age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of major violent events and their relatives, and veteran status.'[296] YouTube's hate speech policy prohibits content if its purpose is to encourage violence or incite hatred against individuals or groups based on these protected attributes.[297] YouTube prohibits threats, including implied ones, and has specific policies against harassment.[298] In rare cases, YouTube may remove content or impose penalties if a creator repeatedly encourages abusive behaviour, targets or insults protected groups, exposes them to physical harm, or harms the platform by inciting hostility for personal gain.[299]

3.3.2 Violent or graphic content policy

---

[290] Pitchbook *Company Profile, YouTube*, available at https://pitchbook.com/profiles/company/51147-55#faqs
[291] YouTube *Community Guidelines: Enforcing Community Guidelines* available at https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#enforcing-community-guidelines
[292] *Ibid.*
[293] *Ibid.*
[294] *Ibid.*
[295] *Ibid.*
[296] YouTube, *Standing Up to Hate*, available at https://www.youtube.com/howyoutubeworks/our-commitments/standing-up-to-hate/
[297] *Ibid.*
[298] *Ibid.*
[299] *Ibid.*

According to YouTube's policy on violent content, any material that is intended to shock or repulse viewers with graphic or gory content or that encourages others to commit acts of violence is prohibited.[300] YouTube prohibits content that incites violence, depicts graphic scenes intended to shock or disgust, such as severe injuries, accidents, or animal abuse and dramatized or fictional footage that lacks clear context.[301] Content such as violent sexual assaults or footage of major violent events recorded by perpetrators is prohibited, even when presented in an educational, documentary, scientific, or artistic context.[302]

3.3.3 Enforcement

3.3.3.1 Content review

YouTube relies on a combination of people and technology to identify and manage inappropriate content.[303] Content could be flagged as inappropriate by automated systems, members of the Priority Flagger program, or users within the broader YouTube community.[304] As of 2022, YouTube had a total of 10,000 content moderators worldwide.[305] To address the challenges of enforcing hate speech and violent or graphic content policies on a global scale, YouTube enhanced its review team by incorporating specialists in linguistics and subject matter.[306] This effort aims to ensure a more nuanced understanding of local languages and contexts.[307] YouTube also makes use of machine learning to improve the detection of potentially hateful content, which is then reviewed by humans.[308] However, the content moderation process is not always applied consistently or objectively. For instance, while YouTube moderators removed videos featuring drug-related violence in Mexico, they allowed violence due to political

---

[300] YouTube *Violent or graphic content policies a*vailable at https://support.google.com/youtube/answer/2802008?hl=en&ref_topic=9282436&sjid=5729834110785531147-EU

[301] *Ibid.*

[302] *Ibid.*

[303] Google Transparency Report *YouTube Community Guidelines enforcement* available at https://transparencyreport.google.com/youtube-policy/removals

[304] *Ibid.*

[305] Legal Resources Centre (LRC) *A Critical Analysis of Content Moderation Policies and the Impact of Spreading Violence, Hatred & Disinformation in the Global South* (2023) pg. 36 available at https://lrc.org.za/wp-content/uploads/LRC-CONTENT-MODERATION-RESEARCH-REPORT.pdf.

[306] *Ibid.*

[307] Google Transparency Report. *YouTube Policy – Hate Speech* available at: https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech

[308] *Ibid.*

conflicts in Syria and Russia to remain on their platform.[309] These inconsistencies were further complicated during the COVID-19 pandemic when workforce reductions forced YouTube to rely heavily on automated systems.[310] As a result, YouTube acknowledged that its increased reliance on automated content removal could lead to the unintended deletion of content that does not violate its policies.[311] Additionally, an investigation by Global Witness and Internet Freedom Foundation revealed that hateful content remained on YouTube despite violating the policies of the platform.[312] Their investigation uncovered posts referring to women as 'dogs,'[313] 'whores'[314], and 'absolute trash.'[315] Of the 79 posts reported to YouTube, the content remained accessible on the platform a month after the content was reported.[316] This demonstrates significant flaws in both human moderation and automated systems, contributing to unfair removals and the persistence of harmful content.

3.3.3.2 Content removal

In terms of YouTube's hate speech or violent and graphic content policy, content that violates these guidelines will be removed.[317] A first-time violation of YouTube's community guidelines will result in a warning without penalties.[318] To prevent further escalation, users with a warning can take policy training to have the warning expire after 90 days, starting from the completion of the training.[319] However, if the same policy is violated within this period, the warning will not expire, and the channel will receive a strike.[320] Violating a different policy after the training will result in a new

---

[309] DeCook, J. R. Cotter, K.,Kanthawala, S. & Foyle, K Safe from "harm": The governance of violence by platforms *Policy and Internet*, (2022) *14*(1), pg. 66 available at https://doi.org/10.1002/poi3.290
[310] Schwemer SF 'Decision Quality and Errors in Content Moderation' (2024) 55(1) *International Review of Intellectual Property and Competition Law pg. 154.*
[311] *Ibid.*
[312] Global Witness *YouTube and Indian Social Media Platform Koo Enable Misogynistic Hate Speech Violating Platform Policies* [Press release]1 February 2024 available at https://www.globalwitness.org/en/press-releases/youtube-and-indian-social-media-platform-koo-enable-misogynistic-hate-speech-violating-platform-policies/
[313] *Ibid.*
[314] *Ibid.*
[315] *Ibid.*
[316] *Ibid.*
[317] Google Transparency Report *YouTube Community Guidelines enforcement a*vailable at https://transparencyreport.google.com/youtube-policy/removals
[318] Google Transparency Report *YouTube Community Guidelines enforcement a*vailable at https://transparencyreport.google.com/youtube-policy/removals
[319] *Ibid.*
[320] Google *Hate Speech Policy Community Guidelines* YouTube Help available at https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436&sjid=5729834110785531147-EU

warning.[321] Accumulating three strikes within 90 days will lead to channel termination.[322] YouTube may terminate a channel or account for repeated violations of Community Guidelines or Terms of Service or after a single instance of severe abuse.[323]

3.3.3.3 Appeals

YouTube was among the first platforms to introduce an appeals process, enabling users to contest strikes for community guidelines violations starting in 2010.[324] YouTube allows content creators to appeal video removals or restrictions.[325] At this stage, a human review of the appeal takes place, where it is determined whether the decision is upheld or overturned.[326]

3.3.3.4 Compliance with Legal Frameworks

YouTube receives content removal requests from various sources, including court orders, national and local government agencies, and law enforcement.[327] Requests can involve single or multiple pieces of content, and sometimes multiple requests may target the same content.[328] YouTube assesses each request for legitimacy and completeness.[329] For a request to be considered, it must be in writing, specify the content to be removed, and clearly explain the legal basis for removal.[330] YouTube reviews court orders to determine their relevance and obligations.[331] Occasionally, YouTube may act on orders not specifically directed at them, respecting court authority.[332] Additionally, if the content has already been removed by the owner or if

---

[321] Google Transparency Report *YouTube Community Guidelines enforcement* available at https://transparencyreport.google.com/youtube-policy/removals
[322] *Ibid.*
[323] Google Transparency Report *YouTube Community Guidelines enforcement* available at https://transparencyreport.google.com/youtube-policy/removals
[324] Diaz A & Hecht-Felella L *Double Standards in Social Media Content Moderation* Brennan Center for Justice, New York University School of Law (2021) pg. 18 available at https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf
[325] Google "Appeals on YouTube" *YouTube Transparency Report*, Available at https://transparencyreport.google.com/youtube-policy/appeals
[326] *Ibid.*
[327] Google. "Government requests to remove content." *Transparency Report*, Available at https://transparencyreport.google.com/government-removals/overview?hl=en
[328] *Ibid.*
[329] *Ibid.*
[330] *Ibid.*
[331] *Ibid.*
[332] *Ibid.*

requests are too vague, YouTube may not take action.[333] Requests to remove content from the entire Internet are also not entertained.[334]

According to YouTube's reporting, between April and June 2024, YouTube removed 3,260,974 channels and 8,497,876 videos in total.[335] Notably, 163,000 of these removals were due to violations of the hate speech policy.[336] Specifically, 764,270 videos were removed from users based in the USA, 103,545 from users based in the UK, and 92,367 from users based in Germany, with no available data for South Africa.[337] This data highlights YouTube's commitment to transparency while also illustrating the scale and scope of its enforcement efforts across different regions.

3.4 X

Whereas YouTube enforces strict community guidelines but makes exceptions for content with educational, documentary, scientific, or artistic value, X's approach is to empower individuals to create and share ideas, information, and opinions without restrictions.[338] Believing that free expression is a fundamental human right, X maintains that everyone should have the opportunity to voice their thoughts.[339] Their role is to facilitate public discourse by ensuring the representation of diverse perspectives.[340] Elon Musk's acquisition of Twitter in October 2022 brought significant changes to content moderation raising concerns about the platform's handling of hate speech.[341] Musk introduced changes to content moderation by reinstating banned accounts and reducing moderation staff.[342] Since Musk's takeover, antisemitic tweets

---

[333] *Ibid.*
[334] *Ibid.*
[335] Google. "YouTube Community Guidelines enforcement" *YouTube Transparency Report*, Available at https://transparencyreport.google.com/youtube-policy/removals
[336] *Ibid.*
[337] *Ibid.*
[338] X *Abuse and Harassment Policy* Help Center, available at https://help.x.com/en/rules-and-policies/abusive-behavior
[339] *Ibid.*
[340] *Ibid.*
[341] Pradel, F., Zilinsky, J., Kosmidis, S. & Theocharis, Y 'Toxic Speech and Limited Demand for Content Moderation on Social Media' (2024) *American Political Science Review* (online publication) pg. 1906 available at https://www.cambridge.org/core/journals/american-political-science-review/article/toxic-speech-and-limited-demand-for-content-moderation-on-social-media/405333D7072585903E81BEF1729378F8
[342] European Union Agency for Fundamental Rights *Online content moderation – Current Challenges in Detecting Hate Speech* (2023) pg. 72 available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2023-online-content-moderation_en.pdf

have increased substantially.[343] The Institute for Strategic Dialogue reported over 325,000 antisemitic tweets from 146,516 accounts between June 2022 and February 2023, with the majority posted after Musk's takeover.[344]

3.4.1 Hateful Conduct Policy

Recognizing that abuse can hinder self-expression, X is dedicated to addressing and preventing abuse driven by hatred, prejudice, or intolerance.[345] X's hateful conduct policy defines hate speech as an attack, dehumanisation, slurs, or hateful imagery based on 'race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.'[346] X's policy prohibits inciting behaviour (including inciting harassment, inciting discrimination, and inciting stereotypes) and reference to violence or violent events, with the primary target being one of the protected categories listed above.[347] This includes spreading harmful stereotypes, promoting harassment, or advocating for discrimination against businesses based on their perceived affiliation with a protected category.[348] X further prohibits targeting individuals with repeated slurs, stereotypes, or content meant to demean or perpetuate negative stereotypes about protected groups.[349] Hateful imagery, including symbols and images that incite hostility, such as hate group symbols, dehumanizing alterations, or references to historical atrocities, is banned from live videos, account bios, and profile/header images and must be marked as sensitive media if posted otherwise.[350] Content that seems hateful in isolation may not be when viewed in context, especially if it involves consensual reclamation of slurs.[351] To properly evaluate such content, X might need to hear from the affected individuals before taking action.[352]

3.4.2 Violent and hateful entities policy

[343] *Ibid.*
[344] *Ibid.*
[345] *Ibid.*
[346] Twitter *Hateful conduct policy*, available at https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.
[347] *Ibid.*
[348] *Ibid.*
[349] *Ibid.*
[350] *Ibid.*
[351] *Ibid.*
[352] *Ibid.*

X further has no tolerance for violent and hateful entities, including terrorist organizations, extremist groups, and individuals who support or are associated with such groups.[353] These entities, through their actions or rhetoric, endanger the safety of others.[354] Users are prohibited from threatening, endorsing, or promoting these groups.[355] Violent entities use physical violence or violent rhetoric against people or infrastructure, while hateful entities systematically encourage or engage in harmful behaviour against protected groups.[356] Content violating this policy includes promoting violence, recruiting for such groups, or distributing related media.[357] Additionally, content intended for educational, documentary, or news purposes is allowed.[358]

3.4.3 Enforcement

3.4.3.1 Content review

X has a global team that enforces its rules, and the team aims to apply the rules objectively and consistently, taking enforcement actions against content that violates these guidelines.[359]

X promotes open discussion of various viewpoints and encourages counter-speech to address misinformation and harmful content.[360] Context is crucial in enforcement decisions, considering factors such as whether content targets specific individuals or groups, the origin of reports, the user's policy history, the severity of the violation, and whether the content is of legitimate public interest.[361] Of all the platforms, X appears to rely the most on artificial intelligence to moderate content.[362] A joint investigation by Global Witness and the Legal Resources Centre tested certain social media platforms by submitting 40 advertisements featuring hate speech targeting women journalists in South Africa, written in English,

---

[353] X. *Violent and hateful entities policy*, Help Center, 2023, Available at https://help.x.com/en/rules-and-policies/violent-entities
[354] *Ibid.*
[355] *Ibid.*
[356] *Ibid.*
[357] *Ibid.*
[358] *Ibid.*
[359] X. *Rules Enforcement*, Transparency Report, Available at https://transparency.x.com/en/reports/rules-enforcement#2021-jul-dec
[360] X, *Our approach to policy development and enforcement philosophy*, Help Center, Available at https://help.x.com/en/rules-and-policies/enforcement-philosophy
[361] *Ibid.*
[362] Legal Resources Centre (LRC) *A Critical Analysis of Content Moderation Policies and the Impact of Spreading Violence, Hatred & Disinformation in the Global South* (2023) pg. 38 available at https://lrc.org.za/wp-content/uploads/LRC-CONTENT-MODERATION-RESEARCH-REPORT.pdf.

Afrikaans, Xhosa, and Zulu.[363] Using advertisements tests the platform's initial defence against hate speech before ad publication, assessing the effectiveness of its content moderation practices.[364] All 40 ads violated X's hate speech policy by labelling women as 'prostitutes, psychopaths, or vermin, and called for them to be beaten and killed.'[365] Despite these ads violating X's policies, all but two were approved for publication.[366] This indicates that AI-driven content moderation is flawed and ineffective if it allows the most obvious hate speech to be approved.[367]

3.4.3.2 Content Removal and Appeal

When enforcement actions are necessary, X may take various steps, including addressing a specific post or direct message, acting against an account, or using both methods.[368] This is done either for violating X's rules or in response to a valid legal request from an authorized entity in a particular country.[369] Enforcement actions can include limiting post visibility, removing posts, adding context labels, or suspending accounts of repeat offenders or those posing risks.[370] When X determines that a post severely violates its rules and needs to be removed, the violator must remove the post before resuming posting content, and they have the option to appeal the content removal request.[371]

3.4.3.3 Compliance with Legal Frameworks

---

[363] Global Witness *Facebook, X/Twitter, YouTube and TikTok approve violent misogynistic hate speech adverts for publication in South Africa* [Press release] 7 December 2023 available at https://www.globalwitness.org/en/press-releases/facebook-xtwitter-youtube-and-tiktok-approve-violent-misogynistic-hate-speech-adverts-publication-south-africa/

[364] Global Witness "*Female stupidity at its best. They all need to die.": Violent and sexualised hate speech targeting women approved for publication by social media platforms* [Press release] 7 December 2023 https://www.globalwitness.org/en/campaigns/digital-threats/south-africa-women-journalists-hate-speech/

[365] *Ibid.*

[366] Global Witness *Facebook, X/Twitter, YouTube and TikTok approve violent misogynistic hate speech adverts for publication in South Africa* [Press release] 7 December 2023 available at https://www.globalwitness.org/en/press-releases/facebook-xtwitter-youtube-and-tiktok-approve-violent-misogynistic-hate-speech-adverts-publication-south-africa/

[367] Global Witness "*Female stupidity at its best. They all need to die.": Violent and sexualised hate speech targeting women approved for publication by social media platforms* [Press release] 7 December 2023 https://www.globalwitness.org/en/campaigns/digital-threats/south-africa-women-journalists-hate-speech/

[368] X, *Our range of enforcement options*, Help Center available at https://help.x.com/en/rules-and-policies/enforcement-options

[369] *Ibid.*

[370] *Ibid.*

[371] *Ibid.*

Many countries have laws that can apply to posts and content on X.[372] To ensure its services remain accessible globally, X may occasionally need to restrict access to certain content in specific countries when receiving valid and properly scoped requests from authorized entities.[373] These restrictions will only affect the jurisdictions issuing the legal demands or where the content violates local laws.[374] Transparency is essential for safeguarding freedom of expression. Accordingly, X has a policy to notify users when their content is withheld.[375] Affected users will be promptly informed within the platform unless a court order under seal prohibits such notification.[376] X is committed to the positive global impact of open information exchange and strives to keep posts flowing freely.[377] The latest global transparency report for the period January to June 2024 indicates that X suspended 5 296 870 accounts and either removed or labelled 10 675 980 posts.[378] Of these actions, 1 105 139 accounts were suspended for hateful conduct, abuse, and harassment, with only 154 of those being automated removals. Additionally, 2,247,107 accounts were removed for hate speech, abuse, and harassment, with just 5,679 of those being automated removals.[379]

3.5 Conclusion

Facebook, YouTube, and X each have comprehensive hate speech and violent content policies that reflect their unique community standards and content moderation approaches. YouTube's definition of hate speech offers the most extensive list of protected characteristics, encompassing additional factors such as age, immigration status, victims of significant violent events and their relatives, veteran status, and gender identity.[380] Similarly, X enforces a broad range of hate speech prohibitions but places significant emphasis on freedom of expression. Facebook distinguishes itself as the only platform that has an independent oversight board that

---

[372] X, *About country withheld content*, Help Center, available at https://help.x.com/en/rules-and-policies/post-withheld-by-country
[373] *Ibid.*
[374] *Ibid.*
[375] *Ibid.*
[376] *Ibid.*
[377] *Ibid.*
[378] X. *Global Transparency Report:H1 2024,* 2024, Available at https://transparency.x.com/content/dam/transparency-twitter/2024/x-global-transparency-report-h1.pdf
[379] *Ibid.*
[380] YouTube *Hate Speech Policy,* YouTube Help available at https://support.google.com/youtube/answer/2801939?ref_topic=9282436#zippy=%

reviews complex or disputed content moderation cases. Neither YouTube nor X has an independent board with similar authority.

While YouTube, Facebook, and X, maintain comprehensive policies to combat hate speech, their consistency and effectiveness in enforcement are often criticized. All policies focus heavily on the rules for users and the consequences of non-compliance, such as warnings, strike systems, suspensions, or permanent account bans. While they emphasize user accountability, they leave open the question of how they hold themselves accountable if their enforcement mechanisms fall short or their policies are violated.[381] Although these enforcement mechanisms seem well-regulated and systemised, they provide social media platforms significant flexibility.[382] Decisions about which content to flag are entirely controlled by the social media platform, and the resulting penalties often reveal a bias.[383] Like many aspects of content moderation, there is limited public information on how these systems operate and their effects on various groups.[384] The self-regulatory practices of social media platforms, which lack accountability for the content they host, raise significant concerns about bias in content removals, errors in enforcement, and the impact on marginalised communities.

[381] Legal Resources Centre (LRC) *A Critical Analysis of Content Moderation Policies and the Impact of Spreading Violence, Hatred & Disinformation in the Global South* (2023) pg. 31 available at https://lrc.org.za/wp-content/uploads/LRC-CONTENT-MODERATION-RESEARCH-REPORT.pdf.
[382] Diaz A & Hecht-Felella L, *Double Standards in Social Media Content Moderation*, Brennan Center for Justice, New York University School of Law (2021) available at https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf
[383] *Ibid.*
[384] *Ibid.*

CHAPTER FOUR
Understanding Global Legal Frameworks

4.1 Introduction

This chapter considers the diverse legal frameworks, policies, and regulations that govern social media platforms and their accountability for hate speech that incites violence. This will be considered by examining four important and relevant jurisdictions: the United Kingdom, the United States of America, Germany, and South Africa.

The examination of these legal frameworks governing social media platforms, will highlight the positive development of the law, as well as highlighting various ongoing challenges and shortcomings. Concerns have been raised by both academics and politicians that self-regulatory frameworks of social media platforms enable private censorship and fail to provide democratic accountability, particularly when user interactions occur without editorial or governmental oversight or when algorithmic decision-making is used.[385] Further, the ongoing development of Internet regulation will be considered within the context of a long-standing debate about the roles and responsibilities of online service providers.[386] This debate, rooted in the structural distinction established in U.S. media policy between carriers and content providers[387] raises important questions about whether these platforms should be viewed as content publishers, thereby bearing certain responsibilities for the material they host, or merely as channels that transmit third party content with minimal accountability.[388]

4.2 International Law

International human rights standards provide a framework that guides national legislation on the regulation of hate speech.[389] Various treaties and conventions outline the responsibilities of both states and companies in preventing and responding to hate speech while promoting respect for human rights. Relevant international instruments outlining the company's duties in terms of hate speech will be detailed below. The Code of Conduct which requires IT companies to

---

[385] Medzini R 'Enhanced Self-Regulation: The Case of Facebook's Content Governance' (2022) 24(10) *New media & society pg. 2228.*
[386] Nash V & Felton L 'Treating the Symptoms or the Disease? Analysing the UK Online Safety Act's Approach to Digital Regulation' (2024) *Policy and internet* pg. 2 https://doi.org/10.1002/poi3.404.
[387] *Ibid.*
[388] *Ibid.*
[389] United Nations Human Rights Council *Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility, or violence* (U.N. Doc. A/HRC/22/17/Add.4), October 2012, pg. 8.

enforce national laws, implement effective processes for reviewing hate speech, educate users on prohibited content, and establish clear community guidelines to address illegal hate speech.[390] Although non-binding, the UNGP, emphasize that businesses must respect human rights, avoid causing adverse impacts, and implement policies and processes to address human rights issues.[391]

Relevant international instruments outlining the State's duties in terms of hate speech will be detailed below. The ECHR emphasizes that exercising freedom of expression carries responsibilities, allowing for necessary legal restrictions in a democratic society.[392] Germany[393] and the UK[394] have signed and ratified the ECHR and are bound by its terms, but the USA and South Africa are not signatories.[395] Additionally, the Council of Europe's Additional Protocol to the Convention on Cybercrime mandates States to criminalize the dissemination of racist and xenophobic material, online threats based on race, ethnicity, or religion,[396] racially motivated insults,[397] distributing content justifying genocide,[398] or assisting in any of the above offenses.[399] The Council of Europe's Additional Protocol to the Convention on Cybercrime is signed and ratified by Germany and signed by South Africa, but neither signed nor ratified by the UK or the USA.[400]

The UNGP, which serves as a guideline and is not binding, asserts that States must safeguard against human rights abuses by third parties, including social media platforms, within their

[390] European Commission *Code of Conduct on Countering Illegal Hate Speech Online* (2016) available at https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.
[391] United Nations High Commissioner for Human Rights *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (2011) available at https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf
[392] Council of Europe, *European Convention on Human Rights,* ETS No. 005, 4 November 1950, Article 10(2).
[393] Council of Europe, *Treaties list for a specific state*, Council of Europe, 2024, Available at https://www.coe.int/en/web/conventions/by-member-states-of-the-council-of-europe?module=treaties-full-list-signature&CodePays=GER&CodeSignatureEnum=&DateStatus=08-09-2024&CodeMatieres=
[394] Dawson J *How might Brexit affect human rights in the UK?* UK Parliament House of Commons Library (7 December 2019) available at https://commonslibrary.parliament.uk/how-might-brexit-affect-human-rights-in-the-uk/
[395] Council of Europe, 'Chart of signatures and ratifications of Treaty 005' available at https://www.coe.int/en/web/conventions/full-list?module=signatures-by-treaty&treatynum=005
[396] Council of Europe, *European Convention on Human Rights,* ETS No. 005, 4 November 1950, Article 4.
[397] Council of Europe, *European Convention on Human Rights,* ETS No. 005, 4 November 1950, Article 5.
[398] Council of Europe, *European Convention on Human Rights,* ETS No. 005, 4 November 1950, Article 6.
[399] Council of Europe, *European Convention on Human Rights,* ETS No. 005, 4 November 1950, Article 7.
[400] Council of Europe, 'Chart of signatures and ratifications of Treaty 189' available at https://www.coe.int/en/web/conventions/full-list?module=signatures-by-treaty&treatynum=189

jurisdiction.[401] This requires the adoption of robust policies, legislation, regulations, and judicial actions to prevent, investigate, penalise, and address such abuses.[402]

Similarly, the 'Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence'[403] (hereinafter "Rabat Plan of Action") also acts as a guide, consolidating the recommendations from multiple expert workshops organized by the OHCHR.[404] These guidelines note that many domestic legal frameworks lack a clear prohibition on incitement to hatred, and those that do use inconsistent terminology.[405] The Rabat Plan of Action further recommends that States should align their domestic laws on incitement to hatred with Article 20(2) of the ICCPR, define key terms like hatred, discrimination, violence, and hostility, and adopt comprehensive anti-discrimination legislation.[406] Article 20(2) of the ICCPR states that 'any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.'[407] Similarly, Article 4 of ICERD requires States to implement legal frameworks against racial discrimination, both at the individual and organizational levels, with necessary consequences.[408] General Recommendation No. 35 also emphasizes that States must adopt legislation to combat hate speech, including criminalising the dissemination of racial hatred, inciting hatred, threatening or inciting violence, as well as justifications for genocide.[409] The UK, USA, Germany, and South Africa have all ratified and are bound by the ICCPR and ICERD[410] which mandate the prohibition of racial hatred that incites violence and require the establishment of appropriate legal frameworks to address such issues. These legal frameworks will be explored in detail below.

---

[401] United Nations High Commissioner for Human Rights, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (2011), pg. 3. Available at: https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf
[402] *Ibid.*
[403] United Nations Human Rights Council *Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility, or violence* (U.N. Doc. A/HRC/22/17/Add.4), October 2012.
[404] United Nations Human Rights Council *Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility, or violence* (U.N. Doc. A/HRC/22/17/Add.4), October 2012.
[405] *Ibid.*
[406] *Ibid.*
[407] International Covenant on Civil and Political Rights, G.A. Res 2200A (XXI) of 16 December 1966 entry into force 23 March 1976, Article 20(2).
[408] *Ibid.*
[409] UN Committee on the Elimination of Racial Discrimination (CERD) *General recommendation No. 35: Combating racist hate speech* 26 September 2013 CERD/C/GC/35.
[410] Office of the United Nations High Commissioner for Human Rights, 'Status of Ratification Interactive Dashboard' available at https://indicators.ohchr.org/.

4.3 United Kingdom ("UK")

As of January 2024, the UK had an active social media audience of 56.2 million users, constituting a considerable 82.8 percent of the population.[411] Given this extensive online presence, safeguarding online users from hate speech becomes paramount. According to UK law, hate speech escalates to an offense or hate crime:- when the perpetrator intends to make the victim fear imminent unlawful violence, seeks to provoke such violence, intends to harass, alarm or distress the victim, or when the victim believes that violence will be used or provoked.[412] Determining the prevalence of how many hate crimes occur online in the UK proves challenging due to insufficient data and inadequate measurement tools.[413] In the most recent Hate Crime statistics from the Home Office, statistics on Online Hate Crime were not provided as there are no comprehensive systems in place to log online hate incidents or crimes.[414] Statistics on Online Hate Crimes were last provided in 2017/2018, when 'experimental'[415] figures were given, which indicated estimated online hate crimes as constituting 2% of all hate crimes recorded in England and Wales.[416] While current and comprehensive data on the extent of online hate in the UK may be lacking, the pervasive influence of hate speech and its impact on real-world events is an ongoing concern.[417] Given these concerns, it begs the question, do social media platforms have any consequences or accountability for hate crimes committed on their platforms in the UK? This research will detail below whether UK legal frameworks hold social media platforms accountable for hate speech that incites violence.

4.3.1 Legal Frameworks and Legislation

4.3.1.1 Domestic Law

On the 26th of October 2023, the Online Safety Act[418] was enacted to make provision for communications offenses.[419] This legislation established a fresh regulatory framework aimed at enhancing the safety of internet services for

[411] Statista, 'Active social media audience in the United Kingdom (UK) in January 2024' (2024) available at https://www.statista.com/statistics/507405/uk-active-social-media-and-mobile-social-media-users/.
[412] *Ibid.*
[413] Stop Hate UK 'What is online hate crime?' Available at https://www.stophateuk.org/about-hate-crime/what-is-online-hate-crime/
[414] *Ibid.*
[415] *Ibid.*
[416] *Ibid.*
[417] *Ibid.*
[418] Online Safety Act 2023, c. 50 (UK).
[419] *Ibid.*

individuals in the UK.[420] The legislation sets forth clear responsibilities for service providers, mandating them to recognize, lessen, and oversee risks stemming from both illegal content and activities, with a particular focus on safeguarding children from harm.[421] In terms of Part 3, Chapter 2 of the Online Safety Act, user-to-user services have a duty of care to ensure their platforms are safe for users.[422] In the context of this Act, a 'user-to-user service'[423] refers to an internet service through which content created directly by a user or uploaded/shared by a user can be accessed by another user or users of the same service.[424] A user-to-user service is defined as a regulated service if it has connections to the UK, such as a substantial user base within the UK.[425] This, by definition, includes the social media platforms Facebook, YouTube, and X. The duties imposed on regulated user-to-user services (including Facebook, YouTube, and X) include the following: illegal content risk assessments,[426] duties regarding illegal content, including reducing the duration of its presence or promptly removing it,[427] content reporting,[428] complying with complaint procedures,[429] freedom of expression and privacy[430] and record-keeping duties.[431]

Furthermore, it endows the regulatory body, the Office of Communications (OFCOM), with enhanced functions and authority.[432] These obligations aim to ensure that online services prioritize safety from the outset.[433] OFCOM has taken on the responsibility of overseeing online safety to ensure that platforms

---

[420] *Ibid.*
[421] Beveridge C 'UK's Online Safety Act 2023: What You Need to Know' *BDO UK* 13 March 2024 available at https://www.bdo.co.uk/en-gb/insights/advisory/risk-and-advisory-services/uks-online-safety-act-2023-what-you-need-to-know.
[422] Online Safety Act 2023, c 50 (UK), Part 2, 3.
[423] *Ibid.*
[424] *Ibid.*
[425] Online Safety Act 2023, c. 50 (UK), Part 2, 4.
[426] Online Safety Act 2023, c. 50 (UK), Section 9.
[427] Online Safety Act 2023, c. 50 (UK), Section 10(2), 10(8).
[428] Online Safety Act 2023, c. 50 (UK), Section 20.
[429] Online Safety Act 2023, c. 50 (UK), Section 21.
[430] Online Safety Act 2023, c. 50 (UK), Section 22(2), 22(3).
[431] *Ibid.*
[432] Online Safety Act 2023, c. 50 (UK), Part 3, 1.
[433] Beveridge C 'UK's Online Safety Act 2023: What You Need to Know' *BDO UK* 13 March 2024 available at https://www.bdo.co.uk/en-gb/insights/advisory/risk-and-advisory-services/uks-online-safety-act-2023-what-you-need-to-know.

prioritize user protection.[434] Following the release of final codes of conduct and guidance by OFCOM, social media platforms will need to demonstrate their compliance with the standards outlined in the Act.[435] The effectiveness of these measures in safeguarding internet users will be monitored by OFCOM, which can enforce action against companies failing to adhere to their duties.[436]

In terms of this Act, OFCOM has the authority to act against any relevant companies, irrespective of their location, if their services have connections to the UK.[437] This encompasses platforms with a sizable user presence in the UK or those specifically targeting UK audiences.[438] Additionally, it applies to services hosting content covered by the law that poses a significant risk of harm to people in the UK.[439] Further, they could examine reports concerning hate speech and the instigation of violence on social media platforms, then implement suitable enforcement measures.[440] Entities could face penalties of up to £18 Million or 10% of the qualifying worldwide revenue, whichever is greater.[441]

The Online Safety Act places the responsibility for determining what constitutes illegal content squarely on the platforms, yet it offers only vague standards for guiding these assessments.[442] The legal requirement that platforms must have 'reasonable grounds to infer'[443] content is illegal means they need to evaluate whether all necessary elements of an offense are present, including mental elements such as intention, while also considering potential defenses.[444] This

---

[434] UK Department for Science, Innovation and Technology *Online Safety Act: Explainer* 92024) available at https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer

[435] *Ibid.*
[436] *Ibid.*
[437] *Ibid.*
[438] *Ibid.*
[439] *Ibid.*
[440] Article 19, *United Kingdom (England and Wales): Responding to 'hate speech'* (2018) available at https://www.article19.org/wp-content/uploads/2018/06/UK-hate-speech_March-2018.pdf
[441] Online Safety Act 2023, c. 50 (UK), Penalties.
[442] Kira B & Schertel Mendes L 'A Primer on the UK Online Safety Act - Key aspects of the new law and its road to implementation' *Verfassungsblog* (13 November 2023) available at https://verfassungsblog.de/a-primer-on-the-uk-online-safety-act/
[443] Online Safety Act 2023, c. 50 (UK), part 11 Section 192(5).
[444] Kira B & Schertel Mendes L 'A Primer on the UK Online Safety Act - Key aspects of the new law and its road to implementation' *Verfassungsblog* (13 November 2023) available at https://verfassungsblog.de/a-primer-on-the-uk-online-safety-act/

creates a challenge for AI-powered automated content moderation systems to effectively conduct such complex legal analysis of content and nuanced assessments.[445] By requiring platforms to make these determinations without clear guidelines, the Online Safety Act amplifies the challenges of ensuring responsible self-regulation and protecting user rights.

4.3.1.2 Illustrative case studies

Since the Online Safety Act is a newly enacted legislation, there are no examples of its application yet. OFCOM, the regulator, hopes to finalise the regulations, guidelines, and codes by 2025, which means the duties of social media platforms in terms of the Online Safety Act are currently non-operational.[446] However, the recent hate speech crisis in Southport has raised concerns about whether the Act is equipped to handle similar situations effectively.[447] After three children were tragically murdered at a dance class in Southport, misinformation spread on social media, falsely accusing a Muslim immigrant of committing the attack.[448] This misinformation evolved into hate speech promoted by far-right activists,[449] which escalated into offline violence.[450] This incident revealed issues with the Online Safety Act and how it would apply to this situation.[451] Firstly, the Online Safety Act addresses incidents not as a single event across platforms but as individual pieces of content.[452]

---

[445] *Ibid.*
[446] Khan S 'Online Safety Act "Not Fit for Purpose" as Far-Right Incites Riots' *The Guardian* 8 August 2024 available at https://www.theguardian.com/media/article/2024/aug/08/online-safety-act-not-fit-for-purpose-far-right-riots-sadiq-khan
[447] Institute for Strategic Dialogue (ISD) 'Southport Riots: An Attempt to Hijack a Really Difficult, Sensitive Issue to Push Hate' *ISD in the Media* 2 August 2024 available at https://www.isdglobal.org/isd-in-the-news/southport-riots-an-attempt-to-hijack-a-really-difficult-sensitive-issue-to-push-hate/
[448] Woods L, Antoniou A & Walsh M *Disinformation and Disorder: The Limits of the Online Safety Act,* Online Safety Act Network (2024) available at https://www.onlinesafetyact.net/analysis/disinformation-and-disorder-the-limits-of-the-online-safety-act/
[449] Khan S 'Online Safety Act "Not Fit for Purpose" as Far-Right Incites Riots' *The Guardian* 8 August 2024 available at https://www.theguardian.com/media/article/2024/aug/08/online-safety-act-not-fit-for-purpose-far-right-riots-sadiq-khan
[450] Institute for Strategic Dialogue (ISD) 'Southport Riots: An Attempt to Hijack a Really Difficult, Sensitive Issue to Push Hate' *ISD in the Media* 2 August 2024 available at https://www.isdglobal.org/isd-in-the-news/southport-riots-an-attempt-to-hijack-a-really-difficult-sensitive-issue-to-push-hate/
[451] Woods L, Antoniou A & Walsh M *Disinformation and Disorder: The Limits of the Online Safety Act,* Online Safety Act Network (2024) available at https://www.onlinesafetyact.net/analysis/disinformation-and-disorder-the-limits-of-the-online-safety-act/
[452] *Ibid.*

This fragmented approach can result in inconsistent responses, with content from the same story being categorized differently, leading to uneven enforcement across platforms.[453] Another concern is the Act's failure to address content that may not appear harmful but becomes harmful through volume and virality.[454] Furthermore, the Online Safety Act overlooks misinformation or disinformation, meaning some harmful posts can evade regulation entirely, limiting its effectiveness in managing a crisis like Southport fuelled by social media.[455] The UK Government suggested the act may be reassessed in response to this incident.[456]

4.4 United States of America (USA)

In comparison, while the UK has implemented and put in place stringent legislation to regulate social media platforms and ensure user protection, the landscape shifts significantly when examining American social media laws. Due to a series of policy decisions, or lack thereof, social media platforms lack accountability.[457] The first significant legislation applying to internet content was the Communications Decency Act of 1996[458] (hereinafter "CDA").[459] The CDA aims to regulate indecency on the Internet.[460] Although many parts of the CDA were struck down by the US Supreme Court in a landmark ruling, namely *Reno v. American Civil Liberties Union*,[461] Section 230, which provides intermediary immunity for hosted content, remained intact.[462] Between 1996 and 2000, US legislation was relatively active, passing several laws to protect children, but these laws frequently faced constitutional challenges that

[453] *Ibid.*
[454] *Ibid.*
[455] *Ibid.*
[456] *Ibid.*
[457] Bayer J, Holznagel B, Korpisaari P & Woods L (eds) *Perspectives on Platform Regulation: Concepts and Models of Social Media Governance Across the Globe* vol 1 (2021) pg. 29 available at doi.org/10.5771/9783748929789
[458] Communications Decency Act 1996, Pub. L. No. 104-104, 110 Stat. 56 (1996).
[459] Bayer J, Holznagel B, Korpisaari P & Woods L (eds) *Perspectives on Platform Regulation: Concepts and Models of Social Media Governance Across the Globe* vol 1 (2021) pg. 13 available at doi.org/10.5771/9783748929789
[460] *Ibid.*
[461] *Reno v. American Civil Liberties Union, 521 U.S. 844 (1997).*
[462] Bayer J, Holznagel B, Korpisaari P & Woods L (eds) *Perspectives on Platform Regulation: Concepts and Models of Social Media Governance Across the Globe* vol 1 (2021) pg. 13 available at doi.org/10.5771/9783748929789

annulled them wholly or in part for violating the First Amendment of the US Constitution, which is detailed and discussed below.[463]

4.4.1 Legal Frameworks and Legislation

4.4.1.1 Domestic Law

The legislation in the USA, unlike the laws in many other liberal democracies, asserts that constitutional safeguards for freedom of speech include public expressions that incite hatred against racial, ethnic, and religious groups.[464] The First Amendment in the US Constitution dictates that 'Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.'[465] The First Amendment of the U.S. Constitution prohibits Congress from enacting laws that limit freedom of speech or the press, thereby shielding social media platforms from government regulation. Even in cases of hate speech, the First Amendment only restricts speech if it incites imminent violence.[466] Balancing national security with free speech rights mandates adjudication, navigating the government's duty to counter terrorism while preserving unrestricted information access and compliance with the Constitution. [467] The Supreme Court's stance against broad content restrictions online, including safeguards for minors, affirms its commitment to upholding the First Amendment's principles. The U.S. Supreme Court has struck down most regulations intended to restrict online content, including those designed to protect minors, citing the constitutional rights in the First Amendment.[468]

---

[463] *Ibid.*

[464] Heyman SJ 'Hate Speech, Public Discourse, and the First Amendment' in Hare I & Weinstein J (eds) *Extreme Speech and Democracy* (2009 pg. 2 available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1186262

[465] Constitution. Amendment I.

[466] Gelashvili T, *Hate Speech on Social Media: Implications of private regulation and governance gaps*. (Master's Thesis, Faculty of Law, Lund University, 2018) pg. 58 available at https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8952399&fileOId=8952403

[467] *Ibid.*

[468] Brown N & Peters J 'Say this not that, Government Regulation and Control of Social Media' (2018) 68 *Syracuse Law Review* pg. 533 available at https://lawreview.syr.edu/wp-content/uploads/2018/10/I-Brown-and-Peters-FINAL-v3.pdf

Not only do social media platforms benefit from the protections of the First Amendment, but they are also generally protected from liability for user-generated content under Section 230 of the CDA.[469] According to Section 230 of the CDA, information provided by users of social media platforms shall not be regarded as having been published or spoken by the social media platforms themselves.[470] Regardless of whether a social media company is aware of inappropriate content or chooses to moderate site content, it generally enjoys immunity from liability for user-generated content.[471] This immunity protection has been upheld and reaffirmed in cases involving claims for defamation, negligence, intentional infliction of emotional distress, and privacy violations.[472] Nevertheless, judicial precedents show that this protection is not absolute or unconditional.[473] Section 230 of the CDA and the protections contained therein are, however, limited under subparagraph (E), which 'excludes protection for federal crimes, communications privacy violations, sex trafficking, or intellectual property infringements.'[474] Over the past two decades, scholars, judges, and legislators have identified various reasons to reconsider the broad immunity provided by Section 230.[475] As the internet has matured, the distinction between online and real-world activities has blurred, and modern platforms operate very differently from the limited services of the 1990s.[476] This immunity perceives platforms as channels for content rather than as publishers, which influences their legal responsibilities. The existing immunity can lead to unfair outcomes, making it difficult for plaintiffs to seek recovery.[477]

---

[469] 47 U.S.C. §230 (1996) Communications Decency Act 1996, Pub. L. No. 104-104, 110 Stat. 56 (1996).
[470] *Ibid.*
[471] Brown N & Peters J 'Say this not that, Government Regulation and Control of Social Media' (2018) 68 *Syracuse Law Review* pg. 538 available at https://lawreview.syr.edu/wp-content/uploads/2018/10/I-Brown-and-Peters-FINAL-v3.pdf
[472] *Ibid.*
[473] Koltay A 'The Protection of Freedom of Expression from Social Media Platforms' (2022) 73(2) *Mercer Law Review* pg. 542 available at https://digitalcommons.law.mercer.edu/cgi/viewcontent.cgi?article=2721&context=jour_mlr
[474] 47 U.S.C. §230 (1996) Communications Decency Act 1996, Pub. L. No. 104-104, 110 Stat. 56 (1996).
[475] McPeak, A **'**Platform Immunity Redefined' (2020) 62(5) *William and Mary Law Review* pg. 1557.
[476] *Ibid.*
[477] *Ibid.*

From an overall perspective, social media platforms have significant protection from liability with the constitutional First Amendment as well as the CDA. However, victims of terrorism have pursued litigation when social media providers have refused to self-regulate by failing to remove many overtly terrorist posts on their own.[478] The basis of their case is that social media platforms aid and provide material support to terrorist organizations by allowing them to use their services for recruitment and propaganda.[479] Under 18 U.S. Code § 2333 - Civil remedies, civil liability arises when an individual aids, abets, or conspires with someone who commits an act of international terrorism by knowingly providing significant assistance.[480] While cases have had mixed results, they highlight ongoing projections to hold platforms accountable for dangerous speech through legal action.[481] This legal avenue of redress against internet content providers who refuse to remove terrorist content is known as the doctrine of material support of terrorism.[482] Incitement and recruitment of terrorists on digital platforms can most effectively be addressed by enforcing this material-support statute.[483] Threatening an organization or providing material support to an organization is an unprotected form of hate speech because it raises grave safety concerns.[484] Should social media platforms be unaware of the harmful posting, they cannot be held accountable.[485] It should, however, be noted that if the social media platform is notified of the offending material by law enforcement agents or private citizens, it may be subject to criminal prosecution for indifference, intransigence, or other failure to act on this information.[486]

4.4.1.2 Illustrative Case Studies

4.4.1.2.1 *Gonzalez v Google*

[478] Tsesis A, *Social Media Accountability for Terrorist Propaganda*, Fordham Law Review Volume 86 Issue 2, 2017 pg. 615 available at https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5444&context=flr.
[479] *Ibid.*
[480] 18 U.S. Code § 2333 - Civil Remedies. Available at https://www.law.cornell.edu/uscode/text/18/2333.
[481] Laub Z *Hate Speech on Social Media: Global Comparisons: CFR Backgrounder* (7 June 2019) available at https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons
[482] Tsesis A 'Social Media Accountability for Terrorist Propaganda' (2017) 86(2) *Fordham Law Review* pg. 615 available at https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5444&context=flr
[483] *Ibid.*
[484] *Ibid.*
[485] *Ibid.*
[486] *Ibid.*

Background

The case centres on Nohemi Gonzalez, a 23-year-old U.S. citizen who was tragically killed during the ISIS attacks in Paris on November 13, 2015.[487] This coordinated series of terrorist acts included shootings at a café, among other locations, and was publicly claimed by ISIS.[488] In response to her death, Reynaldo Gonzalez, Nohemi's father, filed a lawsuit on 14 June 2016 against Google (the owner of YouTube), Twitter, and Facebook for direct and secondary liability for the terrorist attack.[489] On April 21, 2017, Gonzalez filed a Second Amended Complaint (SAC), which added certain family members as Plaintiffs and named Google as the sole defendant.[490] Gonzalez alleged that these platforms facilitated ISIS's operations by allowing them to distribute recruitment and propaganda videos.[491] Gonzalez claimed two of the twelve ISIS terrorists involved in the attacks used social media platforms to share links directing users to ISIS recruitment videos and "jihadi YouTube videos."[492]

Gonzalez Arguments

Gonzalez put forth multiple claims regarding Google's role in facilitating ISIS's use of YouTube. They allege that in terms of § 2333(d) of the Anti-Terrorism Act (ATA), Google is indirectly responsible (Claims One and Two) for aiding and abetting international terrorism and conspiring with ISIS by allowing the terrorist organization to use its platform to spread propaganda and recruit members.[493] Additionally, Gonzalez claims that under § 2333(a) of the ATA, Google is directly accountable (Claims Three and Four) for providing ISIS with material support and resources by enabling its continued presence on YouTube, despite having the capability to remove such content and prevent the re-establishment of banned accounts.[494]

---

[487] *Gonzalez v Google LLC* No. 18-16700 (9th Cir. 2021).
[488] *Ibid.*
[489] *Ibid.*
[490] *Gonzalez v. Google LLC, 2 F 4th 871 (9th Cir. 2021).*
[491] *Gonzalez v Google LLC* No. 18-16700 (9th Cir. 2021).
[492] *Ibid.*
[493] *Gonzalez v. Google LLC, 2 F 4th 871 (9th Cir. 2021).*
[494] *Ibid.*

In terms of 18 U.S.C. § 2333 of the ATA, US nationals are allowed to recover damages for injuries caused by acts of international terrorism.[495] Gonzalez specifically argued that Google (YouTube) aided and abetted international terrorism and enabled the spread of ISIS content by using algorithms to recommend related videos to users[496] based on their viewing history.[497] In this case, Gonzalez asserts that Google algorithms recommended ISIS-related content to users and helped them discover additional ISIS-affiliated videos.[498] By doing so, Gonzalez argues that Google actively assisted ISIS in spreading its propaganda and message.[499] Furthermore, Gonzalez argued that Google (through YouTube) placed paid advertisements near ISIS-related content and shared revenue generated from these ads with ISIS, effectively supporting the group's operations.[500]

Gonzalez further argued that Google was aware of ISIS's use of YouTube to disseminate terrorist propaganda, had received complaints about it, and had the technical ability to remove such content.[501] While Google had at times suspended or blocked certain ISIS-related accounts, Gonzalez contended that these actions were inconsistent and insufficient.[502] Moreover, in certain instances, Google allowed ISIS accounts to remain active if the content did not explicitly violate YouTube's policies or only removed select content while leaving the account operational.[503]

Google's Argument

Google moved to dismiss the claims, asserting protection under Section 230 of the Communications Decency Act (CDA) (47 U.S.C. § 230(c)), which generally shields online platforms from liability for third-party content.[504] Additionally, they requested to dismiss the direct liability claims under § 2333(a), contending

[495] *Ibid.*
[496] *Ibid.*
[497] *Gonzalez v Google LLC* No. 18-16700 (9th Cir. 2021).
[498] *Ibid.*
[499] *Ibid.*
[500] *Gonzalez v. Google LLC, 2 F 4th 871 (9th Cir. 2021).*
[501] *Ibid.*
[502] *Ibid.*
[503] *Ibid.*
[504] *Ibid.*

that these claims did not adequately demonstrate that their actions were the proximate cause of the injuries suffered by the victim.[505]

Application of Legal Frameworks

In *Gonzalez v. Google*, the application of Section 230 of the CDA and 18 U.S. Code § 2333 were central to deciding the case. The district court relied on Section 230 to find that all Gonzalez claims were barred save for the revenue-generating claim, which failed because there was a lack of evidence Google was the proximate cause of the victim's death.[506] The Ninth Circuit subsequently confirmed the district court's ruling, that Section 230 barred the majority of the claims and that the remaining claims of liability lacked adequate legal grounds.[507]

4.4.1.2.2 *Jane Doe v Meta Platforms Inc*[508]

This class action lawsuit was filed by members of the Rohingya community, a Muslim minority group in Myanmar, against Facebook (Meta Platforms, Inc.).[509] The lawsuit included multiple claims regarding Facebook's alleged role in facilitating violent attacks against the Rohingya community.[510]

Background

In 2011, in collaboration with telecommunications companies, Meta pre-installed the Facebook mobile app on new phones sold nationwide in Myanmar to expand internet access.[511] This resulted in 'tens of millions Burmese citizens'[512] having access to Facebook.[513] As social connectivity on Facebook grew, it wasn't long before journalists, academics, and humanitarian leaders alerted Meta that the platform was facilitating the spread of hate speech to incite violence.[514] In August 2017, Myanmar's military launched an ethnic cleansing

[505] *Ibid.*
[506] *Ibid.*
[507] *Gonzalez v Google LLC* No. 18-16700 (9th Cir. 2021).
[508] *Jane Doe v Meta Platforms Inc* No. 21-CIV-06465 (N.D. Cal. 2021)
[509] *Jane Doe v Meta Platforms Inc*, No 21-CIV-06465 (N.D. Cal. 2021), Order Granting Motion to Dismiss.
[510] *Ibid.*
[511] *Ibid.*
[512] *Ibid.*
[513] *Ibid.*
[514] *Ibid.*

campaign against the Rohingya community.[515]A report compiled by Amnesty International in 2022, found that Facebook played a key role in fuelling the violence, with Meta's algorithms amplifying anti-Rohingya content and inciting hatred, which escalated existing discrimination and contributed to the atrocities.[516] After suffering attacks in 2012 and 2017, two members of the Rohingya community brought a class action lawsuit against Meta for the role Facebook played in facilitating these attacks.[517]

Jane Doe's Argument

Jane Doe accepted that their claim would usually be dismissed in U.S. courts due to Section 230 of the CDA but asserted that liability should be assessed under Burmese law, which does not indemnify social media companies from liability for their role in inciting violence.[518]

Meta's Argument

Meta sought to dismiss the Plaintiff's claim as it wasn't brought within the appropriate timelines.[519]

Application of legal frameworks

The claim had to be filed within 2 years of the cause of action, but was filed out of time in 2021.[520] Jane Doe argued that Facebook's role in the violence against the Rohingya was not widely known until 2021. The Court found the claims were still outside the applicable limitations period and dismissed their claims.[521] In March 2024, Jane Doe appealed the decision, and the outcome is still pending.[522] A decision on the merits and Facebook's accountability, will depend

[515]Amnesty International 'The Social Atrocity Meta and the Right to Remedy for the Rohingya' (2022) pg. 6 available at https://www.amnesty.org/en/documents/ASA16/5933/2022/en/
[516]Amnesty International 'The Social Atrocity Meta and the Right to Remedy for the Rohingya' (2022) pg. 7 available at https://www.amnesty.org/en/documents/ASA16/5933/2022/en/
[517] *Jane Doe v Meta Platforms Inc*, No 21-CIV-06465 (N.D. Cal. 2021), Order Granting Motion to Dismiss.
[518] Chander A, 'Section 230 and the International law of Facebook' (2022) 24*Yale Journal of Law & Technology* Pg 407 available at https://law.yale.edu/sites/default/files/area/center/isp/documents/chander.pdf
[519] *Jane Doe v Meta Platforms Inc*, No 21-CIV-06465 (N.D. Cal. 2021), Order Granting Motion to Dismiss.
[520] *Ibid.*
[521]*Ibid.*
[522]*Ibid.*

on the outcome of the appeal and will accordingly only be considered if this appeal is successful.

4.5 Germany

Whereas the USA emphasizes protections under the First Amendment and Section 230 of the CDA, Germany, conversely, implements stringent social media accountability laws that prioritize combating hate speech. This stance is influenced by horrendous incidents of hate speech in their history. The Holocaust is one of the most horrific examples of mass extermination and unconscionable crimes against humanity, where rampant discrimination and hate plagued the European continent.[523] To address the evils of the past, many European nations, including Germany, swiftly enacted hate speech legislation, intending to criminalise speech that targets individuals based on race, ethnicity, religion, and nationality.[524] Germany's hate speech laws are comprehensive and show intention to prevent the spread of hate speech and incitement to violence. Germany's hate speech legislation, dealt with below, forms a robust framework aimed at curbing hate speech and protecting individuals and communities from hate-driven violence. Before 2017 and the implementation of the Network Enforcement Act[525] (NetzDG), the legal framework in Germany included Article 5 of the German Basic Law, which guarantees freedom of expression; however, this right is limited in certain circumstances.[526] Section 130 of the Criminal Code further criminalizes incitement to hatred and attacks on human dignity by insulting, maliciously maligning, or defaming segments of the population. It specifically includes penalties for denying or downplaying the Holocaust. Sections 185-187 of the Criminal Code protect individuals from insults, defamation, and the spread of malicious gossip. The Telemedia Act has specific provisions to protect minors from harmful content, including hate speech. The NetzDG, enacted in 2017, mandates that social media platforms promptly remove illegal content, including hate speech, once it is reported. The NetzDG is likely the first law globally to regulate how social media platforms must handle harmful content on their platforms and the consequences of failing to do so appropriately (Gorwa, 2021).[527]

---

[523] Rauch J 'The Good, the Bad, and the Historically Anti-Semitic: An Analytical Comparison of Anti-Hate Laws in Germany and the United States' (2021) 47(1) *Brooklyn Journal of International Law* pg. 269 available at https://brooklynworks.brooklaw.edu/cgi/viewcontent.cgi?article=1988&context=bjil

[524] *Ibid.*

[525]

[526] Maas S, Wortelker J & Rott A 'Evaluating the Regulation of Social Media: An Empirical Study of the German NetzDG and Facebook' (2024) 48(5) *Telecommunications Policy* Pg 3.

[527] *Ibid.*

4.5.1 Legal Frameworks and Legislation

4.5.1.1 Domestic Law

4.5.1.1.1 Network Enforcement Act (NetzDG)

In 2017, the German Bundestag enacted the NetzDG.[528] The NetzDG applies to Telemedia service providers that aim to make a profit, provide a platform intended to share content with other users, or make the content publicly available and have more than 2,000,000 registered users in Germany.[529] Thus, the NetzDG applies to Facebook, YouTube, and X. The NetzDG mandates social media platforms to review user-reported content and promptly ensure removal should it breach the law.[530] NetzDG allows and empowers social media users and official German bodies to report content they consider unlawful, which must then be removed or blocked from the network.[531]

The NetzDG specifies that only content violating specific enumerated laws must be removed within defined time frames to avoid penalties.[532] Fines for non-compliance are determined by assessing several factors. These include the number of users on each platform, categorized into ranges such as over 20 million, between 4 and 20 million, between 2 and 4 million, and less than 2 million.[533] The seriousness of the offense and any mitigating or aggravating circumstances also play a role in determining the fine.[534] The maximum regulatory fine that can be imposed on legal entities is 50 million Euros.[535]

4.5.1.1.2 Evaluating the Impact of the NetzDG: Benefits and Drawbacks

[528] Network Enforcement Act (NetzDG), 1 September 2017, Bundesgesetzblatt.
[529] *Ibid.*
[530] Kasakowskij T et al 'Network enforcement as denunciation endorsement? A critical study on legal enforcement in social media' (2020) 46 *Telematics and Informatics pg. 1.*
[531] *Ibid.*
[532] *Ibid.*
[533] Bundesamt für Justiz, *Network Enforcement Act Regulatory Fining Guidelines, Guidelines on setting regulatory fines within the scope of the Network Enforcement Act.* Netzwerkdurchsetzungsgesetz - NetzDG 2018 pg. 13 available at chromeextension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.bundesjustizamt.de/SharedDocs/Downloads/DE/NetzDG/Leitlinien_Geldbussen_en.pdf?__blob=publicationFile&v=3
[534] *Ibid.*
[535] *Ibid.*

Numerous scholars, particularly those with legal expertise, have analysed the NetzDG, highlighting its legal flaws and potential impacts. A common concern is that the law may threaten freedom of speech through two main mechanisms: over-blocking (unintentional censorship by third parties) and self-censorship by users.[536] These concerns arise because the innovative nature of the NetzDG leaves its consequences uncertain, potentially leading to these effects on social media platforms.[537] Another significant issue and concern is the privatization of law enforcement, where social media platforms are tasked with determining the legality of content.[538] This is a role traditionally held by legal authorities and with legal acumen.[539] While some argue that social media platforms have always had this responsibility, the consensus is that determining the legality of content is challenging for non-experts.[540] This is a concern that directly parallels issues raised in the UK Online Safety Act. This issue underscores the difficulties of self-regulation in both contexts, highlighting the need for clearer frameworks and standards to ensure accountability and fairness in content moderation.

Despite its shortcomings, the NetzDG is regarded as a ground-breaking regulation for social media content.[541] Internationally, there is ongoing debate about whether similar legislation should be adopted.[542] Some critics argue that the NetzDG could serve as a model for authoritarian regimes to suppress free speech.[543] Conversely, others view the NetzDG as a trailblazer for necessary regulations in other European nations.[544]

Before the NetzDG's implementation, users of Facebook could only report content by clicking a report button next to the offending content.[545] Following the enactment of the NetzDG, Facebook was required to establish a more complex

[536] Maas S, Wortelker J & Rott A 'Evaluating the Regulation of Social Media: An Empirical Study of the German NetzDG and Facebook' (2024) 48(5) *Telecommunications Policy pg. 2.*
[537] *Ibid at pg. 5.*
[538] *Ibid.*
[539] *Ibid.*
[540] *Ibid.*
[541] *Ibid.*
[542] *Ibid.*
[543] *Ibid.*
[544] *Ibid.*
[545] *Ibid at pg. 6.*

mechanism for users to report content that violated the specific laws enumerated in the NetzDG.[546] This new process is more time-consuming and requires users to provide detailed information about themselves and the problem content in question.[547] Additionally, the reporting form for NetzDG violations is only accessible through the imprint sub-section or the Help Centre on Facebook rather than being directly available next to the content.[548] This complexity may have resulted in users continuing to use the simpler, traditional report button to report content under the platform's community standards.[549]

4.5.1.1.2 Digital Services Act (hereinafter "DSA")

In November 2022, a European Union (hereinafter "EU") regulation named the Digital Services Act[550] (hereinafter "DSA") came into force.[551] As it applies to all EU member states from 17 February 2024, this now applies to Germany.[552] The DSA introduces a liability framework that incorporates increased EU oversight and penalties for social media platforms that fail to comply with the obligations in the DSA regulation.[553] These obligations include, among other things, the removal of 'illegal content' such as hate speech, terrorist content, and unlawful discriminatory content, and the obligations of transparency reporting and crisis protocols.[554]

Member States of the EU must ensure that any violations of the obligations under this Regulation are met with appropriate penalties and sanctions.[555] These sanctions should consider the nature, severity, and frequency of the violation, the public interest, the scope of activities, and the economic capacity of the

[546] *Ibid.*
[547] *Ibid.*
[548] *Ibid.*
[549] *Ibid.*
[550] Digital Services Act (EU) 2024.
[551] Department of Enterprise, Trade and Employment, Republic of Ireland *Digital Services Act* available at https://enterprise.gov.ie/en/what-we-do/the-business-environment/digital-single-market/eu-digital-single-market-aspects/digital-services-act/
[552] *Ibid.*
[553] Mchangama J & Alkiviadou N 'South Africa the Model? A Comparative Analysis of Hate Speech Jurisprudence of South Africa and the European Court of Human Rights' (2021) 1 *Journal of Free Speech Law* pg. 545.
[554] European Union, *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services.* L 277/1, 27 October 2022
available at:https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065
[555] *Ibid at pg. L 277/32.*

offender.[556] Penalties should account for systematic or repeated non-compliance, the number of affected service recipients, the intent or negligence behind the infringement, and whether the provider operates in multiple Member States.[557] EU Member States must establish national rules and procedures to ensure fines or penalties are effective, proportionate, and dissuasive for each case, considering all relevant criteria.[558] The maximum fines for non-compliance with obligations can be up to 6% of the provider's annual global turnover from the previous year.[559] The DSA significantly enhances the accountability of social media platforms in handling hate speech that incites violence. By imposing stringent obligations for content moderation, transparency, and risk assessment, coupled with substantial fines for non-compliance, the DSA aims to ensure that platforms take proactive measures to remove illegal content promptly.

Germany's NetzDG holds social media platforms directly accountable for hate speech that incites violence by imposing fines for non-compliance with content removal obligations. Together with the DSA, these regulations create a comprehensive framework that requires social media platforms to take decisive action against hate speech that incites violence. While Germany's approach has made a significant contribution to global discussions on the accountability of social media platforms in content regulation, the NetzDG has faced criticism for its potential to over-block content and the significant burden it places on platforms to evaluate and remove content without having the legal expertise required to make nuanced judgments.

4.5.1.2 Illustrative Case Studies

4.5.1.2.1 NetzDG

In 2019, the Federal Office for Justice, a subdivision of the German justice ministry, fined Facebook 2 Million Euros for failure to comply with the NetzDG.[560] This is the first instance of an American social media platform being

[556] *Ibid.*
[557] *Ibid.*
[558] *Ibid.*
[559] *Ibid at pg. L 277/94.*
[560] Delcker J 'Germany fines Facebook €2M for violating hate speech law' *Politico* 2 July 2019 available at https://www.politico.eu/article/germany-fines-facebook-e2-million-for-violating-hate-speech-law/#:~:text=Europe-

sanctioned for its lack of transparency in managing hate speech.[561] The penalty charge notice specifically criticized Facebook's report for failing to include the number of complaints received regarding unlawful content.[562]

4.5.1.2.2 Digital Services Act

In December 2023, the European Commission initiated the first formal proceeding under the DSA to investigate whether X had violated the DSA concerning its risk management and content moderation practices.[563] This follows a preliminary assessment of X's risk assessment report and transparency initiatives, particularly concerning the dissemination of illegal content linked to Hamas' attacks against Israel.[564] The investigation focused on X's compliance with DSA obligations, including measures to counter unlawful content, the effectiveness of its Community Notes system, transparency in advertising, and the user interface design that may mislead users.[565] In July 2024, X was informed that the European Commission's preliminary finding is that it is in breach of the DSA.[566] The Commission identified the various preliminary non-compliance issues with X.[567] Firstly, verified accounts that contain blue checkmarks mislead and deceive users, as anyone can apply for verification.[568] Secondly, X lacks transparency in advertising, and lastly, X restricts access to public data.[569] If these preliminary findings are confirmed, X would receive a non-compliance decision against X for breaching the DSA.[570] This could result in fines of up to 6% of X's total worldwide annual revenue.[571]

---

[561] *Ibid.*
[562] *Ibid.*
[563] European Commission *Commission opens formal proceedings against X under the Digital Services Act* [Press release] 18 December 2023 available at https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709
[564] *Ibid.*
[565] *Ibid.*
[566] *Ibid.*
[567] *Ibid.*
[568] *Ibid.*
[569] *Ibid.*
[570] *Ibid.*
[571] *Ibid.*

4.6 South Africa

South Africa's hate speech legislation, as in Germany, has been significantly influenced by its history of racist ideology. South Africa's apartheid regime was inherently racist, embedding white supremacy into its laws.[572] The Apartheid government banned speech promoting racial hostility, but used this to suppress anti-apartheid views.[573] The Publications Act of 1974 exemplified this censorship, banning materials deemed obscene, offensive to morals or religion, harmful to state safety, welfare, peace, or public order, disclosing illegal judicial proceedings, damaging group relations, or ridiculing communities.[574] Essentially, this acted as a ban on apartheid hate speech and was used to protect and maintain the apartheid ideal.[575] The end of apartheid 'marked a turning point in South Africa's history,' and South Africa proposed legislation aimed at curbing the dissemination of hate speech and incitement to violence.[576] Detailed below are several laws and legal frameworks in place that are relevant in holding social media platforms accountable for hate speech that incites violence in South Africa.

4.6.1 Legal Frameworks and Legislation

4.6.1.1 Domestic Law

In February 2000, the Equality Act came into force. The purpose of the Act is to give effect to the right to equality to *inter alia* prevent and prohibit hate speech.[577] According to the Equality Act, no person (which includes a juristic and non-juristic entity) may publish, advocate, or communicate any words based on any of the prohibited grounds against another person if such words could reasonably be interpreted as having the intention to 'be hurtful, be harmful or to incite harm, promote or propagate hatred.'[578] If and when appropriate, any matter dealing with the above can be referred to the Director of Public Prosecutions under the common law or other relevant legislation, with

[572] Mchangama J & Alkiviadou N 'South Africa the Model? A Comparative Analysis of Hate Speech Jurisprudence of South Africa and the European Court of Human Rights' (2021) 1 *Journal of Free Speech Law* pg. 560 available at https://futurefreespeech.org//wp-content/uploads/2022/05/Article_South-Africa-the-Model-A-comparative-Analysis-of-Hate-Speech-Jurisprudence-of-South-Africa-and-The-European-Court-of-Human-Rights.pdf
[573] *Ibid.*
[574] *Ibid.*
[575] *Ibid.*
[576] *Ibid.*
[577] Republic of South Africa, Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000, Preamble.
[578] Republic of South Africa, Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000, Section 10.

jurisdiction to institute criminal proceedings.[579] Individuals or organizations can be held accountable for discriminatory practices or hate speech conducted through social media under the Equality Act. However, while the Equality Act imposes obligations on individuals and entities to prevent discrimination and hate speech, it does not explicitly regulate social media platforms.

On the 14th of May 2024, the President assented to the Preventing and Combating of Hate Crimes and Hate Speech (hereinafter "Hate Speech Act").[580] The Hate Speech Act seeks to explicitly criminalize hate speech and hate crimes, providing a clearer framework for holding parties responsible and accountable for failing to address hate speech.[581] The objectives of the Hate Speech Act are to fulfill South Africa's international obligations in addressing prejudice and intolerance, prosecute individuals who commit hate crimes and hate speech with appropriate penalties, and aim to prevent such offenses effectively.[582] It further aims to ensure the enforcement of these measures, facilitate coordinated implementation and administration, and combat hate crimes and hate speech through a unified approach.[583] Additionally, the Act seeks to gather and record data on hate crimes and hate speech, providing a comprehensive framework for addressing these issues within the country.[584] A person who intentionally uses hate speech, propagates hate speech, advocates hate speech, makes available or communicates hate speech to someone with the clear intention of causing harm, inciting harm, or promoting or propagating hatred will be held accountable under the Hate Speech Act.[585] However, this clause, like the Equality Act, does not explicitly impose direct accountability on social media platforms for the content distributed by their users. It criminalizes the actions of individuals who spread hate speech, but does not establish responsibilities or penalties for the platforms themselves. While the Hate

[579] *Ibid.*
[580] Republic of South Africa, Prevention and Combating of Hate Crimes and Hate Speech Act, Act No. 16 of 2023.
[581] *Ibid.*
[582] Republic of South Africa, Prevention and Combating of Hate Crimes and Hate Speech Act, Act No. 16 of 2023, Section 2.
[583] *Ibid.*
[584] *Ibid.*
[585] Republic of South Africa, Prevention and Combating of Hate Crimes and Hate Speech Act, Act No. 16 of 2023, Section 4.

Speech Act and other related laws do not explicitly hold social media platforms accountable in the same way as Germany's NetzDG, they do establish a legal basis for addressing hate speech, which does lead to indirect pressure on social media platforms to accept responsibility to moderate content that incites violence.

4.7 Conclusion

In conclusion, the diverse legal frameworks for social media accountability reflect the evolving landscape of online regulation across different jurisdictions. The UK's Online Safety Act signifies a progressive move towards stringent oversight, aiming to safeguard users by holding platforms accountable for their content and enforcing proactive measures to protect vulnerable groups. As this legislation is newly enacted, only time will tell whether its practical implementation will effectively hold social media platforms accountable.

In contrast to the UK, where social media platforms can be held accountable for incitement to violence, holding platforms accountable in the USA is more complicated, requiring cumbersome and risky legal avenues, such as the material support statute. This is largely due to the protections of the First Amendment, which safeguards free speech, and Section 230 of the CDA, which grants platforms broad immunity for third-party content.

Among the countries studied in this chapter, Germany offers the most legal accountability mechanisms and sets a global benchmark with some of the strictest direct regulations on social media platforms, particularly through the NetzDG. This law mandates the swift removal of illegal content, including hate speech, and imposes significant fines for non-compliance. The combined impact of NetzDG and the Digital Services Act (DSA) compels social media platforms to take decisive actions against hate speech inciting violence, with penalties for non-compliance.

In South Africa, although the Hate Speech Act and related legislation do not explicitly impose accountability on social media platforms akin to Germany's NetzDG, they establish a legal foundation for addressing hate speech. While there is recognition of the need for regulatory oversight in South Africa, the country’s legal approach to social media accountability remains underdeveloped compared to the German and UK models, which

place more specific obligations on platforms to moderate content. A key theme that emerges from this comparative analysis is the tension between self-regulation by social media platforms and the role of governmental oversight. In the UK and Germany, legal provisions place the responsibility on platforms to define illegal content, a self-regulatory approach that can lead to private censorship and reduced accountability. Meanwhile, the U.S. legal framework, particularly Section 230 of the CDA, treats social media platforms as mere channels that host content rather than publishers of such content, resulting in broad immunity that can lead to unfair outcomes. This variation in regulatory approaches underscores the complexities of effectively governing online platforms.

CHAPTER 5

Conclusion

5.1 Summary of Key Findings

The research aimed to consider and assess whether social media platforms are held accountable by global legal frameworks for hate speech that incites violence. Chapter 2 provided a foundational understanding of the definitions of hate speech, exploring how the inconsistency of definitions in both international and domestic legal frameworks complicates the enforcement of hate speech laws across different jurisdictions. It is evident that there is no globally standardised definition of hate speech, resulting in inconsistent application and interpretation by both states and international organizations.[586]

Chapter 3 of the research evaluated the hate speech and violent content policies of three major social media platforms: Facebook, YouTube, and X. The research showed that these platforms have implemented comprehensive hate speech and violent content policies; however, it revealed that their self-regulation often leads to inconsistencies, biases, and negative impacts, especially for marginalized communities. Issues such as algorithmic bias, the difficulty of monitoring diverse languages, and overlooking critical cultural nuances in content moderation proved major obstacles.

The research in Chapter 4 provided a detailed examination of the legal frameworks surrounding social media accountability for hate speech in the UK, USA, Germany, and South Africa. This research highlighted key differences in how these countries address hate speech and their differing and varying levels of accountability for social media platforms. While the UK has progressively introduced new legislative measures to hold platforms accountable, these laws have not yet been tested in their courts. In contrast, in the USA, holding social media platforms accountable for hate speech that incites violence, typically requires litigation under specific laws such as the material support statute. However, as detailed in *Gonzalez v Google*, this is challenging as Section 230 of the CDA, provides platforms with broad immunity from liability for harmful content, even when it incites violence. South Africa's current hate speech law does not hold platforms accountable for hate speech that incites violence, but rather focuses on

[586] Gelashvili T *Hate Speech on Social Media: Implications of private regulation and governance gaps* (Master's Thesis, Faculty of Law, Lund University, (2018) pg. 75 available at https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8952399&fileOId=8952403

individual accountability. Germany's legislation, on the other hand, holds social media platforms directly accountable for failing to comply with the NetzDG by including a penalty provision that imposes fines on platforms that do not meet their obligations.

5.2 Recommendations

To answer the research question, "Do global legal frameworks hold social media platforms accountable for hate speech that incites violence?" it is clear that current legal frameworks are often insufficient. As emphasized in the United Nations Strategy and Plan of Action on Hate Speech, launched in 2019, there is a need for a holistic approach to addressing hate speech globally.[587] It highlights that tackling hate speech is a shared responsibility, requiring action from all sectors of society,[588] which could lead to appropriate accountability.

To ensure greater accountability of social media platforms with hate speech that incites violence on their platform, this holistic approach to tackling hate speech must be adopted. The 2023 UN guide on Countering and Addressing Online Hate Speech (hereinafter "UN guide") offers several key recommendations.[589] Despite ongoing content moderation efforts, social media platforms face significant challenges due to the resource-intensive nature of moderation and the lack of universally accepted standards for managing hate speech.[590] In light of these challenges, the UN Guide suggested social media platforms should be more transparent in their content moderation practices and algorithms,[591] while also considering contextual factors such as history, language, and socio-economic elements.[592] The UN Guide further stresses that platforms must be held accountable to their commitments under the UNGP and ensure appeals, remedial processes, and effective redress channels are accessible and transparent.[593] The UN guide further advocates for stronger judicial mechanisms and independent oversight to ensure

---

[587] United Nations *Countering and Addressing Online Hate Speech: A Guide for Policy Makers and Practitioners* (July 2023) available at: https://www.un.org/en/genocideprevention/documents/publications-and-resources/Countering_Online_Hate_Speech_Guide_policy_makers_practitioners_July_2023.pdf
[588] *Ibid.*
[589] *Ibid.*
[590] Gillespie T, Custodians of the Internet: Platforms, Content Moderation and the Hidden decisions that shape Social Media, (2018) 10(1) *Yale University Press* pg. 9 available at https://www.researchgate.net/profile/Tarleton-Gillespie/publication/327186182_Custodians_of_the_internet_Platforms_content_moderation_and_the_hidden_decisions_that_shape_social_media/links/5dfcffa3a6fdcc2837318e10/Custodians-of-the-Internet-Platforms-Content-Moderation-and-the-Hidden-Decisions-That-Shape-Social-Media.pdf
[591] *Ibid.*
[592] United Nations *Countering and Addressing Online Hate Speech: A Guide for Policy Makers and Practitioners* (July 2023) available at: https://www.un.org/en/genocideprevention/documents/publications-and-resources/Countering_Online_Hate_Speech_Guide_policy_makers_practitioners_July_2023.pdf
[593] *Ibid.*

accountability, urging national courts and independent governance models to review platform practices.[594] Judicial and regulatory mechanisms, including potential reforms of Section 230 of the CDA, could play a crucial role in enhancing the accountability of social media platforms.[595] To improve accountability, international, regional, and domestic legal frameworks need to be harmonized.[596] This would ensure that all platforms are subject to consistent standards across jurisdictions, making it easier to hold them accountable for harmful content.

In conclusion, the research identifies the problems in both the definitions and legal frameworks governing social media accountability for hate speech that incites violence, though there have been notable improvements in legislation and policies to date. Legal frameworks across different jurisdictions are evolving, but they remain inconsistent and often insufficient in compelling platforms to take meaningful action. To effectively hold platforms accountable, a holistic approach is needed, combining the effective implementation of comprehensive content moderation policies with robust domestic judicial mechanisms. This approach will ensure that social media platforms take greater responsibility for moderating harmful content and preventing incitement to violence.

[594] *Ibid.*
[595] Hamilton R J 'Platform-enabled Crimes: Pluralizing Accountability when social media companies enable perpetrators to commit atrocities' (2022) 63(4) *Boston College Law Review* pg. 1414.
[596] Nourooz Pour H 'Transitional Justice and Online Social Platforms: Facebook and the Rohingya Genocide' (2023) 31(2) *International journal of law and information technology* pg. 113.

BIBLIOGRAPHY

PRIMARY SOURCES

Domestic Legislation

Germany

Criminal Code (StGB) Federal Law Gazette I (as last amended by Article 2 of the Act of 22 November 2021) available at https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html.

Network Enforcement Act (NetzDG), 1 September 2017, Bundesgesetzblatt.

Delegated Legislation

Bundesamt für Justiz, *Network Enforcement Act Regulatory Fining Guidelines, Guidelines on setting regulatory fines within the scope of the Network Enforcement Act.* Netzwerkdurchsetzungsgesetz - NetzDG 2018. Available at chromeextension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.bundesjustizamt.de/SharedDocs/Downloads/DE/NetzDG/Leitlinien_Geldbussen_en.pdf?__blob=publicationFile&v=3

Republic of Ireland

Digital Services Act 2024.

Republic of South Africa

- Prevention and Combating of Hate Crimes and Hate Speech Act 15 of 2022.
- Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000.

United Kingdom

- Communications Act 2003.
- Malicious Communications Act 1988.
- Online Safety Act 2023.
- Public Order Act 1986.
- Race Relations Act 1965.

United States of America

- 18 U.S. Code § 2333 - Civil Remedies.
- 47 U.S.C. §230 Communications Decency Act 1996, Pub. L. No. 104-104, 110 Stat. 56 (1996).
- Constitution. Amendment I. Available at https://constitution.congress.gov/constitution/amendment-1/#amendment-1.

International Law

- International Covenant on Civil and Political Rights, G.A. Res 2200A (XXI) of 16 December 1966 entry into force 23 March 1976.
- International Convention on the Elimination of All Forms of Racial Discrimination, GA Res 2106 (XX) of 21 December 1965, entry into force 4 January 1969.
- United Nations General Assembly. *Universal Declaration of Human Rights*, 10 December 1948.
- UN Committee on the Elimination of Racial Discrimination (CERD) *General recommendation No. 35: Combating racist hate speech* 26 September 2013 CERD/C/GC/35.
- United Nations Human Rights Council *Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility, or violence* (U.N. Doc. A/HRC/22/17/Add.4), October 2012

Regional Law

- Council of Europe, *Convention for the Protection of Human Rights and Fundamental Freedoms*, as amended by Protocols Nos. 11 and 14, ETS 5, 4 November 1950.
- Council of Europe, *Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems*, ETS 189, 28 January 2003.
- Council of Europe, *European Convention on Human Rights,* ETS No. 005, 4 November 1950.
- Council of Europe 'Full List: Chart of Signatures and Ratifications by Treaty No. 189' available at: https://www.coe.int/en/web/conventions/full-list?module=signatures-by-treaty&treatynum=189.

- Council of Europe, 'Treaty list for a specific State' available at https://www.coe.int/en/web/conventions/full-list?module=treaties-full-list-signature&CodePays=USA
- Council of Europe, 'Chart of signatures and ratifications of Treaty 005' available at https://www.coe.int/en/web/conventions/full-list?module=signatures-by-treaty&treatynum=005
- European Union, *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services.* L 277/1, 27 October 2022
- Organization of African Unity (OAU), *African Charter on Human and Peoples' Rights* ("Banjul Charter"), 27 June 1981, entry into force 21 October 1986.

Case Law

South Africa

Qwelane v South African Human Rights Commission and Another CCT 13/20 [2021] ZACC 22.

United Kingdom

Connolly v DPP [2007] EWHC 237.

United States of America

- *Brandenburg v. Ohio* 395 U.S. 444 (1969).
- *Reno v. American Civil Liberties Union* 521 U.S. 844 (1997).
- *Gonzalez v Google LLC* 21 F 4th 665 (9th Cir. 2022)
- *Gonzalez v Google LLC* No. 18-16700 (9th Cir. 2021)
- *Reynaldo Gonzalez v Google LLC* 598 US __ (2023)
- *Gonzalez v. Google LLC,* 2 F 4th 871 (9th Cir. 2021)
- *Jane Doe v Meta Platforms Inc* No. 21-CIV-06465 (N.D. Cal. 2021)
- *Jane Doe v. Meta Platforms, Inc.,* No. 21-CIV-06465, *Order Granting Motion to Dismiss*, (N.D. Cal. 2023).

Germany

- *Heike Themel v. Facebook Ireland Inc.* 24.08.2018 - 8 W 1294/18
- Columbia University Global Freedom of Expression, *Heike Themel v. Facebook Ireland Inc.*, available at https://globalfreedomofexpression.columbia.edu/cases/heike-themel-v-facebook-ireland-inc/

SECONDARY SOURCES

Books

Assimakopoulos S, Baide FH, Millar S (eds) *Online Hate Speech in the European Union, a Discourse-Analytic Perspective* (2017) doi.org/10.1007/978-3-319-72604-5

Bayer J, Holznagel B, Korpisaari P & Woods L (eds) *Perspectives on Platform Regulation: Concepts and Models of Social Media Governance Across the Globe* vol 1 (2021) doi.org/10.5771/9783748929789

Keipi T, Näsi M, Oksanen A & Räsänen P *Online Hate and Harmful Content: Cross-National Perspectives* 1 ed (2016).

Wilson P E *The Degradation of Ethics Through the Holocaust* (2023).

Chapters in Books

Heyman SJ 'Hate Speech, Public Discourse, and the First Amendment' in Hare I & Weinstein J (eds) *Extreme Speech and Democracy* (2009).

Khurana U, Vermuelen I, Nalisnick E, Van Noorloos M & Fokkens A 'Hate Speech Criteria: A Modular Approach to Task-Specific Hate Speech Definitions' in Narang K, Davan A M, Mathias L, Vidgen B & Talat Z (eds) *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (2022).

Oberschall A 'Propaganda, hate speech and mass killings' in Oberschall A (ed) *Propaganda, War Crimes Trials and International Law* (2012).

Journal articles

Are C '"Dysfunctional" Appeals and Failures of Algorithmic Justice in Instagram and TikTok Content Moderation' (2024) *Information, communication & society* 1.

Asogwa N & Ezeibe C 'The state, hate speech regulation and sustainable democracy in Africa: a study of Nigeria and Kenya' (2023) 20(3) *African Identities* 199.

Brown N & Peters J 'Say this not that, Government Regulation and Control of Social Media' (2018) 68 *Syracuse Law Review* 521.

Chander A, 'Section 230 and the International law of Facebook' (2022) 24*Yale Journal of Law & Technology* 393.

Chung M & Wihbey J 'Social Media Regulation, Third-Person Effect, and Public Views: A Comparative Study of the United States, the United Kingdom, South Korea, and Mexico' (2022) 26(8) *New media & society* 4534.

De Cook J R, Cotter K, Kanthawala S & Foyle K 'Safe from "Harm": The Governance of Violence by Platforms' (2022)14(1) *Policy & Internet* 63.

Delgado R 'Words that wound A tort action for racial insults, epithets and name calling' (1982) 17 *Harvard Civil Rights-Civil Liberties Law Review* 133.

Etaywe A & Zappavigna M 'The role of social affiliation in incitement: A social semiotic approach to far-right terrorists' incitement to violence' (2024) 53(4) *Language in Society* 623.

Fino A 'Defining Hate Speech: A Seemingly Elusive Task' (2020) 18(1) *Journal of International Criminal Justice* 31.

Gore CD 'The politics of the internet and social media in Africa: three bases of knowledge for advancing research' (2023) 57(1) *Canadian Journal of African Studies / Revue canadienne des études africaines* 201.

Griffin, R 'Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality' (2023) 2(1) *European Law Open* 30.

Hamilton R J 'Platform-enabled Crimes: Pluralizing Accountability when social media companies enable perpetrators to commit atrocities' (2022) 63(4) *Boston College Law Review* 1349.

Hatano A 'Regulating Online Hate Speech through the Prism of Human Rights Law: The Potential of Localised Content Moderation' (2023) 41(1) *The Australian Year Book of International Law Online* 127.

Hietanen M & Eddebo J 'Towards a Definition of Hate Speech—With a Focus on Online Contexts' (2023) 47(4) *Journal of Communication Inquiry* 440.

Jikeli G & Soemer K 'The value of manual annotation in assessing trends of hate speech on social media: was antisemitism on the rise during the tumultuous weeks of Elon Musk's Twitter takeover?' (2023) 6(2) *Journal of Computational Social Science* 943.

Kasakowskij T et al 'Network enforcement as denunciation endorsement? A critical study on legal enforcement in social media' (2020) 46 *Telematics and Informatics* 101317.

Kira B & Schertel Mendes L 'A Primer on the UK Online Safety Act - Key aspects of the new law and its road to implementation' *Verfassungsblog* (13 November 2023) available at https://verfassungsblog.de/a-primer-on-the-uk-online-safety-act/ DOI: 10.59704/2120f79b5f59e60b.

Klonick K 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2020) 129(8) *Yale Law Journal* 2499.

Koltay A 'The Protection of Freedom of Expression from Social Media Platforms' (2022) 73(2) *Mercer Law Review* 523.

Lepoutre M et al 'What Is Hate Speech? The Case for a Corpus Approach' (2023) 18(2) *Criminal Law and Philosophy* 1.

Maas S, Wortelker J & Rott A 'Evaluating the Regulation of Social Media: An Empirical Study of the German NetzDG and Facebook' (2024) 48(5) *Telecommunications Policy* 102719.

MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O 'Hate speech detection: Challenges and solutions' (2019) 14(8) *PLoS one* e0221152.

Mchangama J 'The Sordid Origin of Hate-Speech Laws' (2011) 170 *Policy Review* 45.

Mchangama J & Alkiviadou N 'South Africa the Model? A Comparative Analysis of Hate Speech Jurisprudence of South Africa and the European Court of Human Rights' (2021) 1 *Journal of Free Speech Law* 543.

McPeak A 'Platform Immunity Redefined' (2020) 62 *William and Mary Law Review* 1557.

Medzini R 'Enhanced Self-Regulation: The Case of Facebook's Content Governance' (2022) 24(10) *New media & society* 2227-2251.

Nash V & Felton L 'Treating the Symptoms or the Disease? Analysing the UK Online Safety Act's Approach to Digital Regulation' (2024) *Policy and internet* https://doi.org/10.1002/poi3.404.

Nourooz Pour H 'Transitional Justice and Online Social Platforms: Facebook and the Rohingya Genocide' (2023) 31(2) *International journal of law and information technology*.

Pradel F, Zilinsky J, Kosmidis S & Theocharis, Y 'Toxic speech and limited demand for content moderation on social media' (2024) *American Political Science Review* (online publication) available at https://www.researchgate.net/publication/377658435_Toxic_Speech_and_Limited_Demand_for_Content_Moderation_on_Social_Media DOI:10.1017/S000305542300134X

O'Regan C 'Hate Speech Online: An (Intractable) Contemporary Challenge?' (2018) 71(1) *Current legal problems* 403.

Rauch J 'The Good, the Bad, and the Historically Anti-Semitic: An Analytical Comparison of Anti-Hate Laws in Germany and the United States' (2021) 47(1) *Brooklyn Journal of International Law* 260.

Saurwein F & Spencer-Smith C 'Automated Trouble: The Role of Algorithmic Selection in Harms on Social Media Platforms' (2021) 9(4) *Media and Communication (Lisboa)* 222.

Schwemer SF 'Decision Quality and Errors in Content Moderation' (2024) 55(1) *International Review of Intellectual Property and Competition Law* 139.

Timmermann, KW 'The Relationship between Hate Propaganda and Incitement to Genocide: A New Trend in International Law Towards Criminalization of Hate Propaganda?' (2005) 18(2) *Leiden journal of international law* Pg 257.

Tsesis A 'Social Media Accountability for Terrorist Propaganda' (2017) 86(2) *Fordham Law Review* 605.

Vilar-Lluch S 'Understanding and appraising "hate speech"' (2023) 11(2) *Journal of Language Aggression and Conflict* 279.

Whitten-Woodring J et al 'Poison If You Don't Know How to Use It: Facebook, Democracy, and Human Rights in Myanmar' (2020) 25(3) *International Journal of Press/Politics* 407.

Yanagizawa-Drott D 'Propaganda and Conflict: Evidence from the Rwandan Genocide' (2014) 129(4) *The Quarterly Journal of Economics* 1947.

Zurth P '*The German NetzDG as Role Model or Cautionary Tale? Implications for the Debate on Social Media Liability'* (2021) 31 *Fordham Intellectual Property Media & Entainment Law Journal* 1084.

Reports

Amnesty International *The Social Atrocity Meta and the Right to Remedy for the Rohingya* (2022) available at https://www.amnesty.org/en/documents/ASA16/5933/2022/en/

Article 19 *United Kingdom (England and Wales) Country Report: Responding to 'hate speech'* (2018) available at https://www.article19.org/wp-content/uploads/2018/06/UK-hate-speech_March-2018.pdf

Christchurch Call *The Christchurch Call Story* (15 May 2019) available at https://www.christchurchcall.org/the-christchurch-call-story/

Christchurch Call *Significant progress made on eliminating terrorist content online* (24 September 2019) available at https://www.christchurchcall.org/significant-progress-made-on-eliminating-terrorist-content-online/

Dawson J *How might Brexit affect human rights in the UK?* UK Parliament House of Commons Library (7 December 2019) available at

https://commonslibrary.parliament.uk/how-might-brexit-affect-human-rights-in-the-uk/

Department of Enterprise, Trade and Employment, Republic of Ireland *Digital Services Act* available at https://enterprise.gov.ie/en/what-we-do/the-business-environment/digital-single-market/eu-digital-single-market-aspects/digital-services-act/

Diaz A & Hecht-Felella L *Double Standards in Social Media Content Moderation* Brennan Center for Justice, New York University School of Law (2021) available at https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf

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

European Parliament *Freedom of expression, a comparative-law perspective: The United Kingdom* European Parliamentary Research Service (October 2019) available at https://www.europarl.europa.eu/RegData/etudes/STUD/2019/642263/EPRS_STU(2019)642263_EN.pdf

European Union Agency for Fundamental Rights *Online content moderation – Current Challenges in Detecting Hate Speech* (2023) available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2023-online-content-moderation_en.pdf

French Ministry for Europe and Foreign Affairs The Christchurch Call: What Progress Has Been Made? (12 May 2021) available at https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/the-christchurch-call-what-progress-has-been-made-12-may-2021

Human Rights Watch *Germany: Flawed Social Media Law* (14 February 2018) available at https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law.

Laub Z *Hate Speech on Social Media: Global Comparisons: CFR Backgrounder* (7 June 2019) available at https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

Law Commission *Hate crime laws: Final Report* (2021) Law Com No 402 available at https://assets.publishing.service.gov.uk/media/61ba053ed3bf7f055eb9b8cf/Hate-crime-report-accessible.pdf

Legal Resources Centre (LRC) *A Critical Analysis of Content Moderation Policies and the Impact of Spreading Violence, Hatred & Disinformation in the Global South* (2023) available at https://lrc.org.za/wp-content/uploads/LRC-CONTENT-MODERATION-RESEARCH-REPORT.pdf.

Oversight Board *Homophobic Violence in West Africa, Full Case Decision* (2024) available at https://www.oversightboard.com/decision/fb-ouuwkhko/

Sellars A *Defining Hate Speech* Berkman Klein Center Research Publication No 2016-20, Boston Univ School of Law, Public Law Research Paper No 16-48 (1 December 2016) available at https://ssrn.com/abstract=2882244 or http://dx.doi.org/10.2139/ssrn.2882244.

Southern Africa Litigation Centre and Media Legal Defence Initiative *Freedom of Expression: Litigating Cases of Limitations to the Exercise of Freedom of Speech and Opinion* Litigation Manual Series (2016) available at https://www.southernafricalitigationcentre.org/wp-content/uploads/2017/08/Freedom-of-Expression-Manual.pdf

The Presidency Republic of South Africa *Report of the Expert Panel into the July 2021 Civil Unrest* (29 November 2021) available at https://www.thepresidency.gov.za/sites/default/files/2022-05/Report%20of%20the%20Expert%20Panel%20into%20the%20July%202021%20Civil%20Unrest.pdf

The Wiener Holocaust Library *The Holocaust Explained: The Nazi Rise to Power* available at https://www.theholocaustexplained.org/the-nazi-rise-to-power/the-nazi-rise-to-power/propaganda/

United Nations *Countering and Addressing Online Hate Speech: A Guide for Policy Makers and Practitioners* (July 2023) available at: https://www.un.org/en/genocideprevention/documents/publications-and-resources/Countering_Online_Hate_Speech_Guide_policy_makers_practitioners_July_2023.pdf

UK Department for Science, Innovation and Technology *Online Safety Act: Explainer* 92024) available at https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer

United Nations General Assembly *Minority Issues, Note by the Secretary-General* (16 August 2023) A/78/195.

United Nations General Assembly *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (9 October 2019) A/74/486 available at https://documents-dds-ny.un.org/doc/UNDOC/GEN/N19/308/13/PDF/N1930813.pdf?OpenElement

United Nations High Commissioner for Human Rights *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (2011) available at https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

United Nations Human Rights Council *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (6 April 2018) A/HRC/38/35 available at https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement

United Nations Human Rights Council *Report of the Special Rapporteur on minority issues* (Fernand de Varennes) (3 March 2021) A/HRC/46/57.

University of Oxford, Oxford Pro Bono Publico *Comparative hate speech law: Annexure* Research paper prepared for the Legal Resources Centre, South Africa (March 2012) available at https://www.law.ox.ac.uk/sites/default/files/migrated/1a._comparative_hate_speech_annex.pdf.

Woods L, Antoniou A & Walsh M *Disinformation and Disorder: The Limits of the Online Safety Act,* Online Safety Act Network (2024) available at https://www.onlinesafetyact.net/analysis/disinformation-and-disorder-the-limits-of-the-online-safety-act/

Theses

Gelashvili T *Hate Speech on Social Media: Implications of private regulation and governance gaps* (Master's Thesis, Faculty of Law, Lund University, (2018) available at https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8952399&fileOId=8952403

Conference Proceedings

Olteanu A, Castillo C, De Cristofaro E & Varshney K 'The Effect of Extremist Violence on Hateful Speech Online', paper presented at the *Proceedings of the Twelfth International*

Conference on Web and Social Media (ICWSM) Stanford University California 25- 28 June 2018, available at https://ojs.aaai.org/index.php/ICWSM/issue/view/270

Newspapers

Associated Press 'Myanmar: Facebook accused of letting hate speech thrive' *CBS News* 14 October 2023, available at https://www.cbsnews.com/news/myanmar-facebook-violent-hate-speech-thrives-ap-report/

Beveridge C 'UK's Online Safety Act 2023: What You Need to Know' *BDO UK* 13 March 2024 available at https://www.bdo.co.uk/en-gb/insights/advisory/risk-and-advisory-services/uks-online-safety-act-2023-what-you-need-to-know.

Delcker J 'Germany fines Facebook €2M for violating hate speech law' *Politico* 2 July 2019 available at https://www.politico.eu/article/germany-fines-facebook-e2-million-for-violating-hate-speech-law/#:~:text=Europe-

European Commission *Commission opens formal proceedings against X under the Digital Services Act* [Press release] 18 December 2023 available at https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709

Global Witness *Facebook, X/Twitter, YouTube and TikTok approve violent misogynistic hate speech adverts for publication in South Africa* [Press release] 7 December 2023 available at https://www.globalwitness.org/en/press-releases/facebook-xtwitter-youtube-and-tiktok-approve-violent-misogynistic-hate-speech-adverts-publication-south-africa/

Global Witness *YouTube and Indian Social Media Platform Koo Enable Misogynistic Hate Speech Violating Platform Policies* [Press release]1 February 2024 available at https://www.globalwitness.org/en/press-releases/youtube-and-indian-social-media-platform-koo-enable-misogynistic-hate-speech-violating-platform-policies/

Global Witness "*Female stupidity at its best. They all need to die.": Violent and sexualised hate speech targeting women approved for publication by social media platforms* [Press release] 7 December 2023 https://www.globalwitness.org/en/campaigns/digital-threats/south-africa-women-journalists-hate-speech/

Institute for Strategic Dialogue (ISD) 'Southport Riots: An Attempt to Hijack a Really Difficult, Sensitive Issue to Push Hate' *ISD in the Media* 2 August 2024 available at https://www.isdglobal.org/isd-in-the-news/southport-riots-an-attempt-to-hijack-a-really-difficult-sensitive-issue-to-push-hate/

Karombo T 'South Africa goes after social media as it cracks down on looting and protests' *Quartz Africa* (14 July 2021) available at https://qz.com/africa/2033328/south-africa-to-monitor-social-media-as-protests-rock-the-country

Khan S 'Online Safety Act "Not Fit for Purpose" as Far-Right Incites Riots' *The Guardian* 8 August 2024 available at https://www.theguardian.com/media/article/2024/aug/08/online-safety-act-not-fit-for-purpose-far-right-riots-sadiq-khan

Nxumalo L 'July unrest: Social media fuelled unrest' *IOL* 10 July 2022, available at https://www.iol.co.za/sunday-tribune/news/july-unrest-social-media-fuelled-unrest-481bbafe-1ce1-4ca7-87c9-7da6cfc652eb

Reuters 'Inside Facebook's Myanmar Crisis' *Reuters* 15 August 2018, available at https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

'Britain Passes Sweeping Online Safety Law' *New York Times* 19 September 2023, available at https://www.nytimes.com/2023/09/19/technology/britain-online-safety-law.html

Underwood M 'Is Facebook a Private Company?' *Market Realist* 10 February 2023 available at https://marketrealist.com/p/is-facebook-a-private-company/.

Online Statistics, Social Media and Browsers

European Commission *Commission opens formal proceedings against X under the Digital Services Act* [Press release] 18 December 2023 available at https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709

Facebook *Facebook's Corporate Human Rights Policy*, March 2021, available at https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf.

Facebook *Hate speech policy*, available at https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

Facebook *Understanding Community Standards*, available at https://www.facebook.com/community/using-key-groups-tools/understanding-community-standards/

Google Transparency Report *YouTube Community Guidelines Enforcement,* available at https://transparencyreport.google.com/youtube-policy/removals

Google Transparency Report *YouTube Policy – Hate Speech,* available at: https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech

Google Transparency Report *YouTube Policy Community Guidelines enforcement.* Available at https://transparencyreport.google.com/youtube-policy/removals

Google. "Appeals on YouTube." *YouTube Transparency Report*, Available at https://transparencyreport.google.com/youtube-policy/appeals

Google. "Government requests to remove content." *Transparency Report*, Available at https://transparencyreport.google.com/government-removals/overview?hl=en

Google. "YouTube Community Guidelines enforcement" *YouTube Transparency Report*, Available at https://transparencyreport.google.com/youtube-policy/removals

Google *Hate Speech Policy Community Guidelines* YouTube Help, available at https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436&sjid=5729834110785531147-EU

Meta *Community Standards: Violence and Incitement* Meta Transparency Center, available at https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/

Meta *Community Standards Enforcement Report: Hate Speech on Facebook,* Meta Transparency Center, available at https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook/

Meta *Content Restrictions Report,* Meta Transparency Center, available at https://transparency.meta.com/reports/content-restrictions/

Meta *Counting Strikes,* Meta Transparency Center, available at https://transparency.meta.com/en-gb/enforcement/taking-action/counting-strikes

Meta *Creation of the Oversight Board,* Meta Transparency Center, available at https://transparency.meta.com/en-gb/oversight/creation-of-oversight-board/

Meta *Facebook Community Standards: Hate Speech,* Meta Transparency Center, available at https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/

Meta *How Review Teams Work* Meta Transparency Center, available at https://transparency.meta.com/en-gb/enforcement/detecting-violations/how-review-teams-work/

Meta *How technology detects violations* available at https://transparency.meta.com/en-gb/enforcement/detecting-violations/technology-detects-violations/

Meta *Oversight Board Recommendations,* Meta Transparency Center, available at https://transparency.meta.com/en-gb/oversight/oversight-board-recommendations

Meta *Prioritizing Content Review* Meta Transparency Center, available at https://transparency.meta.com/en-gb/policies/improving/prioritizing-content-review/

Meta *Taking Down Violating Content,* Meta Transparency Center, available at https://transparency.meta.com/en-gb/enforcement/taking-action/taking-down-violating-content/

Office of the United Nations High Commissioner for Human Rights, 'Status of Ratification Interactive Dashboard' available at https://indicators.ohchr.org/.

Office of the United Nations High Commissioner for Human Rights 'Human Rights Indicators' available at: https://indicators.ohchr.org/.

Pitchbook *Company Profile, YouTube*, available at https://pitchbook.com/profiles/company/51147-55#faqs

Statista, 'Active social media audience in the United Kingdom (UK) in January 2024' (2024) available at https://www.statista.com/statistics/507405/uk-active-social-media-and-mobile-social-media-users/.

Stop Hate UK 'What is online hate crime?' available at https://www.stophateuk.org/about-hate-crime/what-is-online-hate-crime/

The Global Statistics 'United Kingdom (UK) Social Media Statistics 2024 Most Popular Platforms' (2024) Available at https://www.theglobalstatistics.com/uk-social-media-usage-statistics/

TikTok *Community guidelines*, available at https://www.tiktok.com/community-guidelines/en/safety-civility/.

Twitter *Hateful conduct policy*, available at https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

World Atlas *Most Popular Social Media Networks in the World* Available at:https://www.worldatlas.com/articles/most-popular-social-media-networks-in-the-world.html

World Population Review *Facebook Users by Country*, available at https://worldpopulationreview.com/country-rankings/facebook-users-by-country

X *Abuse and Harassment Policy* Help Center, available at https://help.x.com/en/rules-and-policies/abusive-behavior

X *Global Transparency Report:H1 2024,* available at https://transparency.x.com/content/dam/transparency-twitter/2024/x-global-transparency-report-h1.pdf

X *Hateful Conduct Policy,* Help Center, 2023, available at https://help.x.com/en/rules-and-policies/hateful-conduct-policy

X *Our approach to policy development and enforcement philosophy*, Help Center, available at https://help.x.com/en/rules-and-policies/enforcement-philosophy

X *Our range of enforcement options*, Help Center, available at https://help.x.com/en/rules-and-policies/enforcement-options

X *Rules Enforcement*, *Transparency Report*, available at https://transparency.x.com/en/reports/rules-enforcement#2021-jul-dec

X *Violent and hateful entities policy*, Help Center, 2023, available at https://help.x.com/en/rules-and-policies/violent-entities

YouTube *Community Guidelines: Enforcing Community Guidelines*, available at https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#enforcing-community-guidelines

YouTube *Content that Violates YouTube's Community Guidelines,* available at https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436&sjid=5729834110785531147-EU

YouTube *Hate Speech Policy,* YouTube Help available at https://support.google.com/youtube/answer/2801939?ref_topic=9282436#zippy=%

YouTube *Standing Up to Hate*, available at https://www.youtube.com/howyoutubeworks/our-commitments/standing-up-to-hate/

YouTube *Understanding YouTube's Community Guidelines Strikes System,* available at https://support.google.com/youtube/answer/9288567?sjid=5729834110785531147-EU

YouTube *Violent or graphic content policies,* available at https://support.google.com/youtube/answer/2802008?hl=en&ref_topic=9282436&sjid=5729834110785531147-EU