Regulating Online Hate Speech And Harmful Content in Aotearoa New Zealand

Beyond Criminalisation And Towards A Statutory Duty of Care

A thesis

submitted in fulfilment

of the requirements for the degree

of

Doctor of Philosophy in Law

at

The University of Waikato

by

RACHEL SUE YIN TAN

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2025

Abstract

This thesis examines how New Zealand regulates online hate speech and harmful content, and evaluates whether current law provides effective protection in the digital environment. The study considers how social-media platforms shape the spread of harmful expression and assesses whether New Zealand's existing legal framework is equipped to respond to these risks while maintaining the right to freedom of expression. The central question guiding the research is whether the present approach is adequate, and what reforms may be needed to address harm more effectively.

The thesis adopts an interpretivist and qualitative methodology, drawing on doctrinal, socio-legal, comparative, and political-legal methods. It uses behavioural, regulatory, and normative theories, including the Online Disinhibition Effect, modalities of regulation, and dignity- and equality-based approaches to free expression. These perspectives help explain why harmful content escalates so quickly online and why traditional legal tools struggle to respond.

The analysis proceeds in three stages. First, it examines the operation of digital platforms, focusing on algorithmic amplification, design choices, and the limits of automated moderation. Second, it reviews New Zealand's legal framework, including the Harmful Digital Communications Act 2015, the Human Rights Act 1993, and the New Zealand Bill of Rights Act 1990. This review shows that the current system is reactive, fragmented, and heavily dependent on voluntary platform policies. Third, the thesis draws comparative insights from the United Kingdom, Australia, France, Germany, and the European Union, where more proactive models, particularly statutory duties of care and transparency obligations, have begun to address platform-level risks.

The research concludes that New Zealand's present approach does not adequately respond to systemic and group-based harms. It argues that a statutory duty of care, supported by risk-assessment requirements, algorithmic transparency, and proportionate safeguards for freedom of expression, offers a more effective and balanced framework. The thesis contributes to existing scholarship by integrating behavioural and regulatory theory with comparative legal analysis and by proposing a model of platform accountability tailored to Aotearoa New Zealand's legal context and human rights commitments.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr Alberto Alvarez-Jiménez, Professor Barry Barton, and Dr Michael Dizon, for their guidance, support, and encouragement throughout this research.

Above all, I would like to thank my wife, Adriana Lorena Barrera Gutierrez, for her unwavering love, patience, and support throughout this journey. Her encouragement, especially during moments of self-doubt, sustained me and made this achievement possible.

I am also deeply grateful to my mother, Catherine Boudville, for her constant support and encouragement, and to my family for their continued support throughout this journey.

This thesis represents the culmination of a challenging and rewarding journey, shaped by growth, resilience, and perseverance.

Table of Contents

Chapter 1: Introduction

Online spaces have become central to how people debate, organise and participate in public life. Hate speech itself is not unique to digital environments and has long existed in offline contexts. Yet these spaces are not equally safe or open. When hateful expression becomes common, people step back. Some speak less; others stay silent. This changes who is heard and who is missing from public debate.

The problem of online hate speech and harmful content becoming common is not only individual behaviour. It is also linked to how digital platforms are built. Algorithms decide what appears, what spreads and what receives attention. They often lift content that creates strong reactions. Sometimes this means conflict or hostility. These design choices allow hate speech and harmful content to proliferate and reach large audiences, making it harder for traditional legal ideas about responsibility and harm to keep pace. Deciding how the law should respond to both user behaviour and platform architecture sits at the centre of this thesis.

1.1 Objectives

Hate speech regulation has become one of the defining tests of how democratic societies govern digital spaces. This thesis aims to critically evaluate the effectiveness of New Zealand's current legal framework in addressing hate speech and harmful content on social media platforms. It investigates whether existing laws adequately protect individuals and communities from harm while upholding the right to freedom of expression and explores how regulatory models from comparable jurisdictions may inform potential reforms.

To achieve this aim, the thesis pursues the following objectives:

1. To examine the scope and limitations of New Zealand's current hate speech laws, with particular reference to the Human Rights Act 1993, the Bill of Rights Act 1990, and the Harmful Digital Communications Act 2015.

2. To assess the legal and regulatory challenges posed by online hate speech, including platform liability, enforcement issues, and to reflect on how the speech and harm balance plays out in digital environments.

3. To conduct a comparative analysis of regulatory frameworks in jurisdictions such as Germany, France, the United Kingdom, and Australia, identifying models that may be adaptable to New Zealand's legal and constitutional context.

4. To analyse the relevance of international human rights law, particularly International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) and General Recommendation No. 35, and evaluate how such instruments inform or influence New Zealand's obligations and domestic responses.

5. To propose a regulatory model for New Zealand that includes a potential statutory duty of care for social media platforms, drawing on lessons from international jurisdictions, while ensuring compatibility with domestic constitutional protections for free speech.

The central research question guiding this thesis is: *How effective is New Zealand's current legal framework in addressing hate speech on social media, and what lessons can be drawn from international approaches to guide potential reforms?*

This research contributes to legal scholarship on digital regulation by situating hate speech at the intersection of constitutional rights, international obligations, and evolving platform governance. It proposes that New Zealand consider adopting a statutory duty of care framework to improve platform accountability, prevent online harm, and promote a safer digital environment. This thesis makes four claims. First, New Zealand's approach to online hate is

reactive and fragmented, leaving platform-level risks under-addressed. Second, behavioural and platform-design dynamics (Online Disinhibition, Social Learning, and engagement-driven architectures) show why criminalisation alone is insufficient. Third, a statutory duty of care, paired with calibrated safe-harbour obligations, offers a proportionate response that advances dignity, substantive equality, and fair participation while remaining consistent with New Zealand Bill of Rights Act 1990 (NZBORA). The New Zealand Bill of Rights Act 1990 provides the constitutional framework within which freedom of expression must be assessed. Its role in shaping proportionality analysis is examined in detail in Chapter 4; it is introduced here to situate the legal discussion within New Zealand's rights-based constitutional context. Fourth, the thesis contributes an integrated framework that links behavioural, regulatory, and normative theory to a New Zealand-specific reform package. The comparative analysis (UK, Australia, France, and Germany) is used not as a blueprint but to identify features that are transferable to New Zealand's legal culture and constitutional settings. The chapters that follow demonstrate how New Zealand can move from reactive moderation to a proactive duty of care framework that is grounded in dignity, equality, and proportionality.

1.2 Background and Context

The regulation of hate speech on social media has become a serious global concern. Online hostility can threaten individuals, communities, and the stability of democratic societies. In New Zealand, online hostility has shaped public debate and raised questions about how well current laws protect communities in digital spaces.

As digital platforms continue to expand, harmful content can spread faster and reach wider audiences. Governments around the world now face a difficult task: protecting people from

harm while still respecting freedom of expression. This thesis examines how New Zealand responds to that challenge and considers how lessons from other countries might guide reform.

Hate speech refers to expression that incites discrimination, hostility, or violence against individuals or groups based on protected characteristics such as race, ethnicity, religion, gender, sexual orientation, or disability.[1] The United Nations defines hate speech as "any kind of communication in speech, writing, or behaviour that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of their identity."[2]

Legal and scholarly definitions vary across jurisdictions, but a commonly cited framework is that of Jeremy Waldron[3], who argues that hate speech undermines the dignity and social assurance of marginalised groups, reinforcing their exclusion from public life.[4] While some legal systems, such as Germany's NetzDG law, adopts a stringent regulatory stance, others, including New Zealand, take a more permissive approach, creating challenges for effective enforcement and content moderation.[5]

This thesis employs a comparative analysis, focusing on New Zealand's hate speech laws alongside approaches of jurisdictions such as the UK, Germany, France, and Australia. Where New Zealand emphasises freedom of expression, Germany and France impose stronger platform duties and content-removal obligations. This research identifies three key gaps in New Zealand's legal approach:

[1] United Nations *Strategy and Plan of Action on Hate Speech*, (2019), at 2
[2] United Nations, above n 1, at 2.
[3] Jeremy Waldron *The Harm in Hate Speech* (Harvard University Press, London, 2012) at 109.
[4] Waldron, above n 3.
[5] David Bromell *Recent Developments in Other Selected Jurisdictions* (Springer Nature, Switzerland, 2022) at 121.

1. Limited platform accountability - unlike Germany's NetzDG law, New Zealand lacks a regulatory framework compelling social media companies to remove harmful content promptly.
2. Ambiguous legal threshold for hate speech - the absence of clear statutory definitions creates inconsistencies in enforcement, leading to uncertainty in legal proceedings.
3. Reactive as opposed to preventive measures - current laws address hate speech only after harm occurs, rather than incorporating proactive regulation, as seen in Australia's Online Safety Act 2021.

Because of these gaps, the present system struggles to reduce online hate effectively. It protects some vulnerable groups but not others, applies vague standards, and depends on weak enforcement. This thesis therefore proposes a hybrid model that balances freedom of speech with stronger, proactive accountability.

The *Harmful Digital Communications Act* 2015 is an important step to regulate online hate speech and harmful content, but it deals mainly with individual distress, not group-based hatred. By contrast, Germany and France hold platforms directly responsible, though their models raise concerns about over-reach. The lack of global consistency means users enjoy uneven protection across jurisdictions.

Globally, the rapid growth of social-media use (now exceeding four billion users) has intensified scrutiny of how platforms amplify harmful content.[6] While platforms like Facebook, X (formerly Twitter) and YouTube facilitate unprecedented connectivity, they also serve as conduits for hate speech, misinformation, and extremist propaganda. This dual function has prompted legislative initiatives such as Germany's NetzDG and the UK's Online

[6] Simon Kemp "Social media users pass the 4 billion mark as global adoption soars" <https://wearesocial.com/cn/blog/2020/10/social-media-users-pass-the-4-billion-mark-as-global-adoption-soars/>

Safety Act. This thesis addresses these gaps by evaluating New Zealand's current legal framework and examining whether reforms are necessary to ensure its ongoing relevance.

By engaging in a comprehensive analysis of these issues, this research contributes to the growing body of knowledge on the intersection of law, technology, and social media's societal impacts. Its findings will offer valuable insights for policymakers, lawmakers, and other stakeholders in developing robust strategies that promote a safer and more inclusive digital environment.

In New Zealand, survey data by Netsafe (2019) showed that about 15 percent of adults experienced online hate speech in the previous year, up from 11 percent the year before.[7] This increase shows how online hostility affects public discourse and wellbeing.[8] Repeated exposure to hateful content can cause anxiety and disengagement. These effects call for a move away from purely reactive laws.

Online hate erodes trust, discourages civic participation, and can entrench social exclusion. The negative impact of online hate speech on digital discourse and societal cohesion is profound. As hate speech proliferates within virtual environments, it erodes the normative foundations of democratic engagement, fostering a climate of antagonism and exclusion.[9] The pervasive dissemination of discriminatory rhetoric not only diminishes trust within online communities, but also contributes to the epistemic silencing of marginalized groups, effectively constraining their capacity to engage in public discourse.[10] The enforcement of content moderation policies, coupled with the implicit threat of content removal or account

[7] Angela Boundy "2019 online hate speech insights" (12 December 2019) Netsafe - Online Safety Help and Advice for New Zealanders <https://www.netsafe.org.nz/2019-online-hate-speech-insights/>.
[8] Boundy, above n 7.
[9] Waldron, above n 3.
[10] Katharine Gelber and Luke McNamara "Evidencing the harms of hate speech" (2015) Social Identities 324 at 335.

suspension, compels users to self-censor their speech to avoid penalties. Such suppression alters online discourse, as individuals, particularly those from marginalized communities, moderate their expressions out of fear of algorithmic enforcement, reducing the diversity of perspectives in digital spaces.[11]

Beyond its digital manifestation**,** the effects of online hate speech extend offline too. They **can** exacerbate intergroup tensions and reinforce systemic discrimination.[12] Empirical studies indicate that prolonged exposure to online hate speech can erode civic engagement, induce psychological distress, and, in extreme cases, contribute to offline hostility and violence. Hawdon et al. highlight that individuals who frequently encounter hate speech online are more likely to experience heightened anxiety, social withdrawal, and disengagement from political or civic participation, as these environments become perceived as hostile or unwelcoming**.**[13] These factors necessitate a reassessment of regulatory approaches, particularly in jurisdictions where hate speech legislation remains reactive rather than pre-emptive.[14]

Addressing this issue necessitates an examination of the underlying mechanisms that facilitate disseminating hateful expression on social media platforms. It demands a comprehensive analysis of the regulatory landscape governing online communication and the efficacy of current policies in combatting hate speech.[15]

A clear understanding of how social media, hate speech and harm interact is essential for finding effective solutions. Through an interdisciplinary approach that amalgamates legal,

[11] Tarleton Gillespie "Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media" (2018) Yale University Press at 177.
[12] Gillespie, above n 11 at 182.
[13] James Hawdon, Atte Oksanen, and Pekka Räsänen "Exposure to Online Hate in Four Nations: A Cross-National Consideration" (2017) 38 Deviant Behavior 254 at 265.
[14] Bromell, above n 5, at 121.
[15] Nicole Stremlau and Iginio Gagliardone "Socio-legal approaches to online hate speech" in Naomi Creutzfeldt, Marc Mason, and Kirsten McConnachie (eds) *Routledge Handbook of Socio-Legal Theory and Methods* (Routledge, 2019) 385 at 245.

sociological, psychological, and ethical perspectives, the academic community can take significant strides towards nurturing a safer and more inclusive digital environment wherein individuals can engage in free expression without fear of malicious persecution or harm.

The events of 15 March 2019 intensified national concern about online hate speech.[16] Netsafe[17] received approximately 600 complaints and inquiries directly linked to hateful expressions following the Christchurch attack, showing how quickly online hostility can escalate and how difficult it is for reactive laws to respond.[18]

In the context of New Zealand, survey shows that online hate is increasingly widespread, with a significant share of adults reporting exposure or harm.[19] Public concern about the proliferation of hateful content remains high and many New Zealanders believe platforms should bear greater responsibility for addressing it.[20]

Legal frameworks governing online hate speech must account for the fluid, decentralized, and culturally embedded nature of digital discourse. As Stremlau & Gagliardone argue, online hate speech presents unique socio-legal challenges, as traditional legal interventions often struggle to regulate dynamic and rapidly evolving expressions of harm.[21] These complexities raise significant questions about the effectiveness of existing legal mechanisms, particularly in jurisdictions that lack clear statutory definitions, enforceability measures, or a balance between platform accountability and free speech protections. A socio-legal approach to regulation is

[16] Neil Melhuish and Edgar Pacheco “Measuring trends in online hate speech victimisation and exposure, and attitudes in New Zealand” (December 12, 2019) Social Science Research Network, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3501977> at 1.
[17] Netsafe "What is the HDCA?" (2021) <https://netsafe.org.nz/what-is-the-hdca/>. Netsafe is currently designated to fulfil the tasks and functions of the Approved Agency pursuant to the Harmful Digital Communications Act 2015.
[18] Melhuish and Pacheco, above n 16, at 1.
[19] Melhuish and Pacheco, above n 16, at 1.
[20] Melhuish and Pacheco, above n 16, at 1.
[21] Nicole Stremlau and Iginio Gagliardone "Socio-legal approaches to online hate speech" in Naomi Creutzfeldt, Marc Mason, and Kirsten McConnachie (eds) Routledge Handbook of Socio-Legal Theory and Methods (Routledge, 2019) 385.

therefore essential, as it considers not only the legal classification of hate speech but also the broader political, technological, and social conditions that influence its spread and impact. In addition, a comprehensive understanding of the psychological and sociological factors that influence the creation and spread of online hate speech is required.[22] Scholars may investigate the motivations, biases, and group dynamics that lead individuals and communities to express hatred. Understanding these underlying processes can illuminate potential strategies for promoting empathy, tolerance, and digital citizenship.

Legal analysis is needed to assess how current frameworks work and where reform is necessary. In addition, fostering digital media literacy and promoting ethical engagement in the digital domain emerge as essential components for the development of resilient and empathetic online communities.

In New Zealand, this idea finds partial reflection in the Harmful Digital Communications Act 2015[23], which enables individuals to file complaints through NetSafe[24], the approved agency tasked with addressing online harm. While this mechanism focuses primarily on harmful digital communications rather than hate speech per se, it represents a legally endorsed form of user-initiated regulation, reinforcing the notion that end-users can act as first responders in identifying harmful content.

Following Christchurch, the government reviewed whether existing laws were adequate.[25] The *Human Rights Act* 1993 prohibits incitement of racial disharmony, yet it does not extend comparable protection for religion, sexual orientation, and gender identity.[26] In 2019, then-

[22] Stremlau and Gagliardone above n 15, at 245.
[23] Harmful Digital Communications Act 2015, s 8.
[24] Netsafe "Making a Complaint" <https://www.netsafe.org.nz/complaint/>
[25] Derek Cheng "Christchurch Call update: Social media giants join forces to fight extremism" *New Zealand Herald* (online ed, 23 September 2019).
[26] Human Rights Act 1993, s 61.

Justice Minister Andrew Little announced a wider review led by the Ministry of Justice and the Human Rights Commission.[27] Proposed reforms to expand hate-speech offences met strong political resistance and were shelved. The Law Commission has since released its report which recommends strengthening the *Human Rights Act* by adding new protected grounds, including gender identity and innate variations of sex characteristics.[28] Broader recommendations on hate speech and incitement are expected to continue shaping the policy debate.

Because social-media communication is borderless, any response must also reflect New Zealand's international human-rights obligations. The country has ratified the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), including General Recommendation No. 35 on hate speech. [29] These instruments are not directly binding in domestic law, but courts often use them to interpret local statutes. Sections 61 and 131 of the Human Rights Act 1993 criminalise incitement of racial disharmony but remain narrower than what ICERD recommends. Section 14 of the New Zealand Bill of Rights Act 1990 guarantees freedom of expression, so any new restriction must be proportionate and justified.

1.3 Methodological Approach

This research is grounded in an interpretivist paradigm, which assumes that legal meaning and regulatory effectiveness are shaped by social, institutional, and cultural contexts rather than existing as fixed, objective truths.[30] It adopts a constructivist ontology, recognising that

[27] Question for Written Answer 29678 (15 August 2019) (Tim Macindoe).
[28] Law Commission Hate Crime Law Reform (2024) <https://www.lawcom.govt.nz/our-work/hate-crime/> and Law Commission "Ia Tangata, A review of the protections in the Human Rights Act 1993 for people who are transgender, people who are non-binary and people with innate variations of sex characteristics" (2025) <https://www.lawcom.govt.nz/our-work/ia-tangata/tab/report>
[29] International Convention on the Elimination of All Forms of Racial Discrimination *General Recommendation No. 35 Combating Racist Hate Speec*h CERD/C/GC/35 (26 September 2013) at [19-20].
[30] Michael Crotty *The Foundations of Social Research: Meaning and Perspective in the Research Process* (Sage Publications, London, 1998) at 66-67.

concepts such as “harm,” “speech,” and “accountability” acquire meaning through interpretation by lawmakers, courts, and platform actors. Within this framework, law is understood not merely as a set of rules but as a dynamic process influenced by social norms and institutional behaviour. This thesis uses a multi-method legal research approach, combining doctrinal, socio-legal, comparative, and political-legal methodologies to provide a comprehensive analysis of New Zealand’s current legal framework on online hate speech and potential reforms.

The research adopts a qualitative doctrinal and comparative legal methodology. The qualitative dimension refers to interpretive analysis of legal texts, regulatory frameworks, and policy materials rather than empirical measurement. The thesis evaluates how different legal systems conceptualise harm, responsibility, and proportionality in responding to online speech, enabling comparative assessment of regulatory design across jurisdictions. While this research mainly uses doctrinal and comparative methods, it also follows a clear qualitative process. Primary legal materials were chosen because they relate directly to the research question. These include statutes, parliamentary debates, government and regulatory reports, and case law. Relevant secondary sources such as journal articles and commentaries were considered. The approach is interpretive and mainly deductive, guided by legal meaning and context rather than numerical data. The countries for comparison were purposely selected because they show different legal systems but also have relevance to New Zealand’s situation. This way, the analysis stays interpretive and context-based instead of aiming for statistical generalisation.

Doctrinal legal research forms the foundation of this thesis, enabling a critical examination of New Zealand’s existing legal framework and its capacity to address online hate speech effectively. This involves analysing the provisions and limitations of key legislation, such as

the Harmful Digital Communications Act 2015, the Human Rights Act 1993, and the Bill of Rights Act 1990. By focusing on the interpretation and application of these laws, doctrinal analysis provides the groundwork for identifying gaps and inconsistencies in the current legal landscape.

The United States is not included as a primary comparative jurisdiction in this study. While the United States has generated extensive scholarship on freedom of expression and hate speech, the strong constitutional protection afforded by the First Amendment produces a regulatory environment that differs significantly from the statutory frameworks examined in this thesis. Because the thesis focuses on regulatory models that impose platform obligations or statutory duties of care, the U.S. approach provides limited guidance for reforms within New Zealand's constitutional setting.

The comparative approach investigates how other jurisdictions have addressed online hate speech, focusing on the United Kingdom, Australia, France, and Germany. These countries were selected for their diverse legal traditions and regulatory frameworks, offering valuable insights for New Zealand: United Kingdom: The UK's common law system shares historical and structural similarities with New Zealand, making its legislative approaches particularly relevant; Australia: As a neighbouring country with similar legal and cultural contexts, Australia provides a regional comparison and insights into harmonisation opportunities; France and Germany: Both countries have enacted robust regulations (e.g., France's Avia Law and Germany's NetzDG) to address online hate speech, showcasing lessons from civil law systems and highlighting the balance between regulation and freedom of expression. This comparative analysis will identify successes and challenges in these jurisdictions, offering potential pathways for reform tailored to New Zealand's context.

The thesis also uses thematic analysis to organise and interpret legal and policy materials. In practice, this means paying attention to recurring concepts such as harm, platform responsibility, freedom of expression and substantive equality.

This thesis also draws on political-legal methodology to understand how political discourse and societal values shape legislative responses to online hate speech and harmful content in New Zealand. This perspective helps situate the legal framework within broader debates about free speech and state responsibility.[31]

The thesis situates its analysis within a broader human rights framework, focusing on the balance between freedom of expression and protection from harm. Relevant international human rights instruments include the International Covenant on Civil and Political Rights (ICCPR) and the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), both of which inform New Zealand's approach to freedom of expression and hate speech regulation. Drawing on Article 19 of the International Covenant on Civil and Political Rights (ICCPR), this framework evaluates the legitimacy of restrictions on freedom of expression to prevent hate speech. These international principles guide the examination of New Zealand's domestic laws, including the Harmful Digital Communications Act 2015 (HDCA), Human Rights Acts 1993, Summary Offences 1981, Bill of Rights Act 1990, Broadcasting Act 1989, Film, Video, Publications Classification Act 1993, and Crimes Act 1961. The analysis also draws on comparative governance approaches, such as those

[31] Maxim Institute Podcast "4. David Hall and Alex Penk on hate speech and free speech - Maxim Institute Podcast - Podcast" Podtail <https://soundcloud.com/user-864022290/4-david-hall-and-alex-penk-on-hate-speech-and-free-speech>.

developed within the European Union, to understand how supranational bodies address cross-border challenges posed by social media platforms.

1.4 Scope and Qualifications of the Thesis

This research has several limitations. First, it focuses on selected jurisdictions for comparison. The United Kingdom, Australia, France and Germany were chosen because their regulatory approaches offer lessons for New Zealand, but they do not represent all global models. Their relevance depends on constitutional and cultural differences that limit direct transferability.

Second, the thesis draws on publicly available legal and policy material. It does not include interviews, surveys or direct engagement with platform regulators or affected communities. While this approach allows for a clear doctrinal and socio-legal analysis, it cannot fully capture the lived experience of online harm or the internal decision-making processes of companies.

Third, the regulatory environment is changing quickly. New laws such as the EU Digital Services Act and UK Online Safety Act continue to evolve, and platform design and moderation systems also shift over time. The thesis reflects the most up-to-date information available at the time of writing, but some developments may move beyond the scope of this study.

Finally, the thesis is limited to hate speech and harmful content. Hate speech constitutes the thesis's primary analytical focus, examined as a distinct form of group-based harm raising particular concerns for dignity, equality, and democratic participation. The broader category of harmful content is used descriptively to situate hate speech within the wider regulatory

environment of platform-mediated harms, recognising that online harms often operate along a continuum rather than within discrete legal categories. The thesis therefore analyses harmful content primarily insofar as it illuminates the systemic dynamics through which hate speech is amplified and governed online. This conceptual distinction informs the terminology and analytical structure adopted throughout the thesis. It does not address broader online harms such as privacy breaches, cyberbullying outside the Harmful Digital Communications Act 2015, or emerging issues relating to generative AI. These areas offer opportunities for future research.

1.5 Significance of the Thesis

This research matters for several reasons. First, it brings together legal, behavioural, and technological perspectives to explain why online hate speech is difficult to regulate within New Zealand's current framework. By combining doctrinal, social-legal and comparative approaches, the thesis offers a more complete picture of the problem than studies that focus on only one area.

Second, the research shows how design choices made by social-media platforms shape the spread of harmful content. This links legal questions about responsibility to the realities of how digital systems work. Understanding this connection is essential for developing regulation that is practical as well as principled.

Third, the study identifies gaps in New Zealand's current laws and explains why they remain reactive and fragmented. By comparing international models, it thus highlights options that

could be adapted to New Zealand's constitutional setting, while also showing the limits of direct transplantation.

Finally, the thesis makes a forward-looking contribution by proposing a statutory duty of care tailored to New Zealand. This model provides a structured way to strengthened platform accountability while still protecting freedom of expression. This proposal is intended to support policymakers, courts and regulators as they consider how best to address online harm in the coming years. These contributions show why a new regulatory model is needed and how this thesis provides a pathway for it. This thesis therefore approaches online hate speech as a systemic regulatory problem situated within broader platform-mediated harms, while maintaining hate speech as its central analytical focus.

1.6 Main Conclusions of the Thesis

This thesis concludes that New Zealand's current approach is reactive, fragmented, and overly reliant on self-regulation by social-media platforms. Legal remedies focus on individual harm rather than systemic risk, leaving gaps in protection for marginalised communities. By contrast, international developments show that proactive frameworks (this is especially of those based on statutory duties of care, transparency, and risk assessment) offer more effective ways to prevent harm before it occurs.

The research contributes to the literature in four key ways. First, it develops an integrated theoretical framework that links behavioural theories (such as the Online Disinhibition Effect and Social Learning Theory), regulatory theories (including Lessig's modalities of regulation and Murray's dynamic regulation), and normative theories (from Waldron's dignity-based

approach, Fredman's substantive equality, and Breyer's proportionality model). This synthesis provides a new conceptual lens for analysing online hate speech in Aotearoa New Zealand. Second, it advances a socio-legal understanding of how legal, political and cultural factors shape the country's cautious approach to hate speech regulation, demonstrating that New Zealand's emphasis on freedom of expression reflects deeper traditions of legal and political culture rather than simply legislative inertia.

Third, it contributes original comparative analysis by identifying transferable features from the UK, Australian, French and German frameworks while highlighting the limits of transposability due to New Zealand's distinct constitutional and institutional setting.

Finally, it offers a policy contribution by proposing a statutory duty of care model bespoke to New Zealand; supported by algorithmic transparency, procedural safeguards and co-regulatory mechanisms. This model provides a forward-looking blueprint for embedding dignity, equality and democratic participation into the governance of digital platforms.

1.7 Overview of thesis

The following chapters build on these findings and develop the analysis in a structured way. This thesis investigates whether New Zealand's current legal framework is adequately equipped to regulate online hate speech and harmful content, particularly on social-media platforms. It asks whether existing laws strike a fair balance between freedom of expression and protection from harm, and whether a statutory duty of care could provide a more effective and proportionate response. Drawing on behavioural, regulatory, and normative theories, the thesis builds a foundation for evaluating New Zealand's laws, comparing international models,

and proposing reforms that promote accountability, dignity, and democratic participation in the digital environment. These objectives flow through the structure of the thesis, with each chapter addressing a different aspect of the research question.

Chapter 1 introduces the context and objectives of the study. It explains the rise of online hate speech and harmful content globally and within New Zealand, and the challenges this poses for regulators seeking to protect individuals and communities while preserving free expression. The chapter outlines the limitations of the current framework, identifies key research gaps, and sets out the mixed-method approach combining doctrinal, socio-legal, comparative, and political-legal analysis.

Chapter 2 develops the theoretical foundation that underpins the thesis. It introduces behavioural theories such as the Online Disinhibition Effect and Social Learning Theory to explain how online environments encourage harmful communication. It also examines regulatory theories, including Lessig's modalities of regulation and Murray's model of dynamic regulation, which help to conceptualise the interaction between law, technology, and social norms. Finally, it considers normative theories of harm, dignity, and equality advanced by Waldron, Fredman, and Breyer. Together, these perspectives provide the conceptual framework for evaluating how law can prevent and respond to online hate speech and harmful content.

Chapter 3 turns to the operation of digital platforms themselves. It analyses how algorithms, platform design, and amplification mechanisms contribute to the spread of hateful and harmful material. The chapter discusses definitional and technical challenges in identifying online hate speech, as well as the limits of automated moderation and detection. It argues that regulatory

responses must account for these systemic features of digital communication, since they shape the visibility and persistence of harmful expression.

Chapter 4 examines New Zealand's domestic legal framework. It analyses the Human Rights Act 1993, the Harmful Digital Communications Act 2015, and the New Zealand Bill of Rights Act 1990, as well as related statutes such as the Broadcasting Act 1989 and the Crimes Act 1961. The chapter evaluates how these laws address, or fail to address, online hate speech and harmful content. It concludes that while existing legislation provides partial remedies, it remains reactive, fragmented, and limited in scope. The analysis highlights the absence of consistent definitions, platform accountability, and enforcement capacity.

Chapter 5 provides a comparative analysis of international responses. It reviews developments in the United Kingdom, Australia, France, Germany, and the European Union, focusing on frameworks such as the UK Online Safety Act 2023 and the EU Digital Services Act 2022. These models demonstrate how jurisdictions have moved towards proactive regulation through platform duties of care, transparency obligations, and content-removal requirements. The comparison shows that New Zealand lags behind in adopting structural reforms capable of preventing harm before it occurs.

Chapter 6 builds on these findings to propose a tailored regulatory model for New Zealand. It evaluates the feasibility of introducing a statutory duty of care for social-media platforms, drawing on the UK and Australian approaches. The chapter argues that such a framework could enhance accountability while remaining compatible with New Zealand's constitutional commitment to free expression. It also considers how public-private partnerships, co-regulatory mechanisms, and user-empowerment strategies could complement legal reform.

Chapter 7 concludes the thesis by integrating the theoretical, legal, and comparative insights. It finds that New Zealand's current framework does not adequately respond to the scale or complexity of online hate speech and harmful content. A proactive, duty-of-care-based model is recommended to bridge the gap between protection from harm and respect for freedom of expression. The chapter emphasises that reform must be grounded in human-rights principles, proportionality, and substantive equality, ensuring that New Zealand's digital spaces remain both safe and open.

Chapter 2: Theoretical Framework

This chapter develops the theoretical framework that underpins the thesis. It brings together behavioural, regulatory, and normative perspectives to explain why online hate speech and harmful content emerges, how it can be addressed, and why principled regulation is necessary. As outlined in Chapter 1, online hate speech presents a complex challenge at the intersection of law, technology, human rights, and digital platform design. It shapes who can participate in public discourse and raises difficult questions about harm, accountability, and equality in the digital age.

This chapter introduces three sets of theories. First, behavioural and psychological theories, including the Online Disinhibition Effect and Social Learning Theory, explain how online platforms amplify harmful speech by lowering social restraints and rewarding hostility. Second, regulatory theories, particularly Lessig's four modalities and the work of Reidenberg and Murray, highlight the interaction between law, norms, markets, and technological design in shaping digital communication, and reveal the limits of criminal law alone. Third, normative theories of speech and equality, drawing on Waldron's dignity-based approach, Fredman's

substantive equality model, and related perspectives, explain why hate speech is not only offensive but socially harmful in ways that justify legal intervention.

Together, these theories provide a structured framework for analysing the adequacy of New Zealand's current legal response. They also guide the thesis's central claim: that criminalisation alone is insufficient, and that a regulatory model grounded in behavioural insight, responsive regulation, and substantive equality may better address the nature of online harm.

2.1 Literature Review and Theoretical Debates

This chapter provides the conceptual and theoretical foundations for the thesis. It integrates insights from behavioural science, legal theory and human rights scholarship. It begins by reviewing key literature on the nature and regulation of online harm, particularly hate speech on social media, before introducing the theoretical models that underpin this study.

The literature reflects three dominant currents of debates:

- Behavioural theories examine the psychological and environmental conditions that give rise to online harm. Suler's work on the online disinhibition effect reveals how anonymity and invisibility alter user behaviour in digital contexts, often lowering thresholds for aggression and hostility.[36] Studies by Lapidot-Lefler and Barak and Wachs and Wright, further confirm the link between disinhibition and the frequency of hate speech in anonymous environments.[37]

[36] John Suler, "The Online Disinhibition Effect" (2005) 2(2) International Journal of Applied Psychoanalytic Studies 184 at 184-188.
[37] Noam Lapidot-Lefler and Azy Barak "Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition" (2012) 28 Computers in human behavior at 435 ;Sebastian Wachs and Michelle F Wright "Associations between Bystanders and Perpetrators of Online Hate: The Moderating Role of Toxic Online Disinhibition" (2018) 15 International Journal of Environmental Research and Public Health at 1.

Social learning models, especially those influenced by Bandura and Sutherland, suggest that online hate is acquired through exposure to aggressive norms and peer reinforcement.[38] The viral structure of social platforms can serve as a vector for socialisation into extremist or prejudicial attitudes.[39]

- Regulatory and governance theories address how legal and technological frameworks structure user behaviour. Lawrence Lessig's four modalities of regulation - law, norms, market and code- remain foundational in evaluating digital accountability mechanisms.[40] Murray expands on these principles in his model of dynamic regulation, calling for adaptive, multi-factor responses to the complexities of the online environment.[41]
- Finally, the normative legal theorists such as Waldron argues that hate speech causes not only emotional distress but also corrodes the public good of mutual assurance.[42] His framework distinguishes between offence and harm, insisting that regulation must focus on the impairment of dignity and social inclusion. Mills' classical liberal argument, that speech should only be restricted where it causes harm to others, provides a philosophical starting point for these debates.[43] In more recent work, Fredman calls for a substantive equality approach that situates speech within the social structures that reinforce discrimination.[44]

The chapter proceeds as follows: Section 2.2 analyses behavioural theories that explain why online hate speech and harm emerge; Section 2.3 examines regulatory approaches that address how these behaviours are shaped by law, markets, norms, and code; Section 2.4 investigates

[38] Tim Boone and others "Social Learning Theory Albert Bandura Englewood Cliffs, N.J.: Prentice-Hall, 1977. 247 pp. paperbound Group & Organisation Studies, at 384.
[39] Paul Benjamin Lowry and others "Why Do Adults Engage in Cyberbullying on Social Media? An Integration of Online Disinhibition and Deindividuation Effects with the Social Structure and Social Learning Model" 2016 27(4) Information Systems Research.
[40] Lawrence Lessig "The New Chicago School" (1998) 27 The Journal of Legal Studies 661 at 664.
[41] Andrew Murray *The Regulation of Cyberspace* (Routledge-Cavendish, Abingdon, 2007) at 240.
[42] Jeremy Waldron *The Harm in Hate Speech* (Harvard University Press, London, 2012) at 5.
[43] John Stuart Mill *On Liberty*, edited by David Bromwich, and George Kateb (Yale University Press, 2003) at 80.
[44] Sandra Fredman "Substantive Equality Revisited" (2016) 14(3) International Journal of Constitutional Law 712 at 735.

normative theories of free expression, harm, and substantive equality; Section 2.5 synthesises these perspectives and offer concluding reflections.

These frameworks will together lay the groundwork for the thesis's central inquiry: How effective is New Zealand's current legal framework in addressing online hate speech and harmful content on social media, and what lessons can be drawn from international approaches to guide potential reforms? By integrating insights from behavioural psychology, regulatory theory and normative legal philosophy, the chapter equips the thesis to critically evaluate the adequacy of domestic law and the scope for adopting more effective, rights-consistent models drawn from comparative jurisdictions.

2.2 Behavioural Theories of Online Harm

The Online Disinhibition Effect (ODE), introduced by psychologist John Suler, provides a foundational behavioural theory for understanding why individuals may engage in hostile or extreme expression online, particularly on social media. Suler classifies this effect into two broad forms: benign disinhibition, involving increased openness or emotional candour, and toxic disinhibition, encompassing hostile, aggressive, or discriminatory behaviours such as harassment, trolling, and hate speech. He attributes these behaviours to six interrelated features of digital environments:

i.) dissociative anonymity - that allows users to separate their actions from their real-world identity, reducing personal accountability;

ii.) invisibility - this removes visual cues and feedback, weakening self-regulation and empathy;

iii.) asynchronicity, - delays responses and disrupts the social cues that typically moderate behaviour.

iv.) solipsistic introjection - occurs when individuals internalise online interlocutors, imagining them as part of their own psyche, which lowers interpersonal boundaries.

v.) dissociative imagination - lets users mentally separate online actions from real-world consequences, treating the digital space as a fictional or game-like arena.

vi.) attenuated status and authority - minimises perceived hierarchy, encouraging users to speak or act more freely than they would in offline interactions.[45]

Each of which contributes to a psychological distancing from the consequences of one's speech.[46] For example, invisibility reduces the role of non-verbal cues and emotional feedback, while anonymity shields the speaker's identity, undermining accountability. These features interact to weaken normative social restraints, leading individuals to express themselves in ways they might suppress in physical or regulated public settings. Suler distinguishes between benign disinhibition, which involves openness and candour, and toxic disinhibition, which encompasses hostile behaviours such as harassment, trolling, or hate speech. This thesis focuses on toxic disinhibition, as it most directly explains how platform features facilitate harmful online conduct.[47]

For this thesis, toxic disinhibition is especially relevant because it illustrates how harmful online behaviour is not simply a matter of individual malice, but a product of psychological and structural factors embedded in digital platforms. Suler's theory thus supports a broader analytical claim: platform features are not neutral, but play an active role in shaping user conduct.[48] This aligns with the argument that any effective legal response to online hate in New

[45] John Suler, "The Online Disinhibition Effect" (2005) 2(2) *International Journal of Applied Psychoanalytic Studies* 184 at 184-188.
[46]John R. Suler *Psychology of the Digital Age: Humans Become Electric* (Cambridge University Press, Cambridge, 2015) at 97 and 102.
[47] Suler, above n 46, at 102.
[48] John Suler, "The Online Disinhibition Effect" (2005) 2(2) *International Journal of Applied Psychoanalytic Studies* 184 at 188.

Zealand should not rely solely on criminalisation. Instead, a regulatory approach grounded in a statutory duty of care would more directly address the platform-level conditions that facilitate and amplify toxic disinhibition.

To accurately frame toxic disinhibition in relation to regulatory concerns, it is necessary to distinguish between aggression, harassment, and hate speech, as well as the conditions under which they arise online. These terms, while often used interchangeably in public debate, represent different behavioural patterns with distinct implications. Toxic disinhibition is not simply a matter of users "being rude online"; rather, it reflects a deeper shift in social behaviour enabled by the psychological and technological environment of digital platforms. According to Suler, toxic disinhibition emerges when users feel disconnected from the real-world consequences of their words and actions.[49] This disconnection lowers emotional self-regulation, reduces empathy, and can lead to impulsive or extreme expression.

As Lapidot-Lefler and Barak demonstrate in their experimental research, features such as anonymity, invisibility, and lack of eye contact play a key role in weakening social restraints and heightening online hostility. [50] Their findings support the notion of *online disinhibition* which is a psychological state in which users feel immune from consequences. This is a core component of toxic disinhibition.[51] The online environment, particularly when it includes such features, removes the normal cues that would encourage reflective or prosocial communication.

Toxic disinhibition is thus a psychological condition in which users experience reduced self-awareness, reduced accountability, and increased emotional arousal; a combination that

[49] Suler, above n 46, at 97.
[50] Lapidot-Lefler and Barak, above n 37, at 441.
[51] Lapidot-Lefler and Barak, above n 37, at 442.

increases the likelihood of harmful behaviour. This harmful behaviour often manifests as aggression. When users experience the disinhibiting effects of anonymity, invisibility and emotional detachment, they are more prone to act on hostile impulses. In this context, aggression becomes a natural extension of toxic disinhibition rather than a separate or unrelated phenomenon. The user's lowered self-regulation and empathy create the ideal conditions for verbal attacks or threats to emerge, particularly in environments where social norms are weak or absent. As such, toxic disinhibition can be understood not only as a precondition but also as a psychological mechanism that increases the likelihood of aggressive expression online.

Aggression refers generally to hostile or harmful actions, which can be impulsive or reactive, and may manifest online as verbal attacks or threats[52]. Harassment involves repetitive, targeted actions aimed at demeaning or intimidating an individual. In contrast, hate speech is primarily defined by its focus on group-based hostility, targeting people based on protected characteristics such as race, religion, gender, or sexual orientation.[53]

These behaviours may occur independently or in combination. For instance, hate speech may be delivered in an aggressive manner, or sustained as a form of harassment. However, they are not interchangeable and do not operate in a fixed sequence. Hate speech, for example, does not always involve aggression or harassment. It may appear in a single statement that promotes group-based hostility without being personally targeted or emotionally charged. For example, a social media post that claims a minority group is "dangerous" or "inferior", even if made once and without abusive language, may still qualify as hate speech due to its discriminatory intent and potential to incite prejudice. In such cases, the harm lies not in aggression or

[52] Lapidot-Lefler and Barak, above n 37, at 436.
[53] Stephanie Tom Tong "Social Processes of Online Hate" in Diana E Forsythe and Ashley Marie Mehlenbacher (eds) Digital Hate: The Global Convergence of Hate Online (Routledge, London, 2023) 24, at 46.

repetition, but in the propagation of harmful stereotypes and the normalisation of exclusionary ideologies.

Building on this, Wachs and Wright show that toxic disinhibition not only facilitates harmful behaviour, but also influences the social roles that users assume online. Their study highlights how bystanders, when disinhibited, may shift from passive observers to active participants in online hate.[54] This reinforces the idea that online harm is not only an individual psychological issue, but a socially learned and platform-mediated phenomenon. The term *socially learned* refers to how harmful behaviours are modelled, reinforced, and normalised through digital group interactions, while *platform-mediated* highlights the way social media architecture (including algorithms, moderation practices, and interface design) can amplify or suppress these behaviours. These insights lend weight to the argument that New Zealand's current legal framework must go beyond punishing individual perpetrators. A viable regulatory response must address the design and governance of social media platforms themselves. These two dimensions, social learning and platform architecture, are explored further in Chapters 3 and 6. There, the thesis examines how harmful norms are reinforced through online group dynamics and how platform features like algorithms and interface design play a central role in enabling or deterring online hate.

As social media and its users proliferate, so too would the notion of keyboard warriors[55]. This is a term described as individuals who engages in online discourse, often around politically charged or identity-based issues. These users typically advocate strongly for their views,

[54] Sebastian Wachs and Michelle F. Wright "Associations between Bystanders and Perpetrators of Online Hate: The Moderating Role of Toxic Online Disinhibition" (2018) 15 International Journal of Environmental Research and Public Health at 3.
[55] Graham Edward Geddes *Keyboard Warriors : The Production of Islamophobic Identity and an Extreme Worldview within an Online Political Community* (Cambridge Scholars Publishing, Newcastle-upon-Tyne, United Kingdom, 2016).

positioning themselves as defenders of truth or national values, and expressing heightened concern about controversial issues such as immigration, religious diversity or hate speech regulation. While not necessarily always engaging in hate speech, their discursive strategies often involve generalisation, moral polarisation and othering, particularly against minority groups. [56] This therefore falls under the purview of being toxic disinhibition and this theory will be the primary focus and referred to mainly because online hate speech does resonate under the negative category. As discussed, toxic disinhibition operates alongside social learning mechanisms and platform architecture to normalise harmful behaviour online.

Geddes observes that keyboard warriors commonly frame themselves as champions of free speech while expressing concern that legal responses to online hate (such as criminalisation or expanded protections for minority groups) may amount to state censorship or overreach.[57] This creates a discursive paradox: while acknowledging the problem of hate speech, keyboard warriors often resist regulatory intervention, fearing erosion of their expressive freedoms.

This paradox underscores the importance of thoroughly engaging with the concept of freedom of expression, which is examined in detail in Chapter 3, particularly in relation to its legal boundaries and tensions with hate speech regulation. Their activity illustrates how online platforms foster a particular mode of identity performance that blends ideological intensity with toxic disinhibition. Although these individuals may behave moderately in offline settings, the affordance of online anonymity, asynchronous interaction and platform amplification contribute to a more aggressive, oppositional communicative style. Offline, the same individuals may not display overt hostility or extremism, often maintaining socially acceptable

[56] Geddes, above n 55, at 33-35.
[57] Geddes, above n 55, at 109.

conduct in face-to-face environments. The disconnect between offline moderation and online aggression is part of what makes the keyboard warrior archetype notable. Their heightened online expression is not necessarily a sign of persistent hate speech behaviour, but rather an indicator of how certain users become emboldened by the digital environment. As Geddes notes, this behaviour reflects not only individual attitudes but also the affordances of online spaces that reward performative outrage. [58] Keyboard warriors do not always engage in hate speech per se, but the probability increases when their discourse centres on sensitive socio-political issues, especially when group-based identity is involved. The term "warrior" metaphorically captures their combative style and self-ascribed mission to defend contested values, often leading to inflammatory or exclusionary speech that can border on or cross into hate speech, depending on context and content.

This behavioural archetype is significant for this thesis because it exemplifies how hate-adjacent speech is normalised, repeated and often given visibility through platform dynamics. The keyboard warrior represents not only an individual actor shaped by online disinhibition, but also a symptom of structural conditions that reward confrontation and penalise nuance. As such, regulatory frameworks focused solely on individual intent or isolated acts of hate speech risk overlooking the cultural and technological ecosystem in which these figures thrive. This reinforces the thesis' argument for a statutory duty of care, aimed at addressing the systematic conditions under which harmful speech escalates and circulates online. This proposed regulatory approach is explored in depth in Chapter 6, where the duty of care framework is analysed in relation to platform design, content moderation, and legal accountability. The next section introduces this concept and explains why current legal responses in New Zealand fall short of addressing the structural dimensions of online harm.

[58] Geddes, above n 55, at 109.

Lapidot-Lefler and Barak identify that anonymity plays a contributing factor to toxic online disinhibition, but their analysis goes beyond simply asserting that people behave badly when unnamed.[59] Anonymity, in psychological terms, reduces self-awareness and dissolves the social cues that normally regulate behaviour in face-to-face interactions. Drawing on deindividuation theory, Lapidot-Lefler and Barak argue that when users perceive their actions as untraceable or disconnected from their real-world identities, their behaviour becomes less constrained by internalised social norms.[60] Anonymity removes both external accountability and internal self-monitoring, creating conditions where users feel less responsible for the social impact of their speech.[61] However, anonymity alone does not lead universally to antisocial behaviour or hate speech. As their study acknowledges, the effects of anonymity are contingent on contextual factors, including group norms, individual personality traits, and perceived audience expectations.[62] This helps explain why not all anonymous users engage in harmful speech; but under certain conditions, anonymity increases the likelihood of such behaviour by lowering perceived accountability. Each of these contextual factors (group norms, individual personality traits, and audience expectations) plays a crucial role in shaping how anonymity manifests in online interactions.

Group norms determine the behavioural baseline within specific online communities; in environments where hostility or bigotry is normalised, anonymous users are more likely to mimic such behaviour.[63] These norms often emerge through repeated interactions, influential

[59] Lapidot-Lefler and Barak, above n 37, at 435.
[60] Anonymous, 'To Reveal or Not to Reveal: A Theoretical Model of Anonymous Communication' (1998) 8(4) Communication Theory 381 at 385.
[61] Lapidot-Lefler and Barak, above n 37, at 436.
[62] Lapidot-Lefler and Barak, above n 37, at 436.
[63] Leonie Rösner and Nicole C. Krämer "Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments". Social Media + Society, 2(3) <https://doi.org/10.1177/2056305116664220>

users modelling particular speech styles and reinforcement mechanisms such as shares, likes and algorithmic promotion. For example, on fringe platforms such as 4chan or certain Reddit threads, users frequently post inflammatory or derogatory content which is then upvoted or echoed by other users. Over time, this creates a localised culture where such speech becomes expected (even encouraged), thereby shaping how new users behave and what types of expression are considered acceptable.

Personality traits, such as impulsivity or low empathy, further modulate how anonymity reduces self-regulation, aligning with findings by Lapidot-Lefler and Barak.[64] Finally, perceived audience (how users imagine or interpret the presence of others) is highlighted by Lea et. al. in their Social Identity Model of Deindividuation Effects (SIDE), which shows that under anonymity, people shift from personal to group-based identity and behaviour.[65] Lea et. al. argue, anonymity does not inherently produce antisocial behaviour; rather, it amplifies the influence of salient group norms, making context and audience perception critical factors in shaping online expression[66] Likewise, the perceived nature of the audience, that is, whether imagined as sympathetic, hostile, or indifferent, affects how users calibrate their tone and content. These psychological and social variables interact with technological features to influence whether anonymity leads to toxic or benign forms of expression.

Lapidot-Lefler and Barak also make an important distinction between anonymity and invisibility, two concepts that are often conflated but operate differently. Anonymity relates to the absence of identity markers, whereas invisibility refers to the lack of visual feedback in online interaction, that is, users cannot see each other or their reactions. Invisibility further

[64] Lapidot-Lefler and Barak, above n 37, at 436.
[65] Martin Lea, Russell Spears, Daphne de Groot "Knowing Me, Knowing You: Anonymity Effects on Social Identity Processes within Groups." Personality & Social Psychology Bulletin 27, no. 5 (2001): 526–37 at 536.
[66] Lea, Spears and de Groot, above n 65, at 536.

weakens empathy and self-regulation by obscuring the human consequences of one's words.[67] Unlike anonymity, which might still be present in pseudonymous or closed-group contexts, invisibility is almost universal in online environments and plays a profound role in flattening affective responses. The absence of facial expressions (that is, from behind screens), body language, or even delayed written reactions contributes to a sense of emotional detachment, which facilitates toxic behaviours. Their findings also underscore that invisibility may neutralise social identity cues such as race, gender, disability, or physical appearance; making prejudiced users feel free to project stereotypes without challenge.[68] Invisibility thus not only contributes to the disinhibition of aggression but also makes it easier for users to objectify or dehumanise others, especially when these others are constructed abstractly through speech or avatars. This adds another layer to toxic disinhibition: it is not merely a loss of self-restraint, but a shift as to how users perceive online interaction itself. Others are no longer seen as real individuals with feelings and rights, but as abstract or dehumanised targets.

These insights reinforce the need to understand online hate speech and harmful content not just as individual deviance, but as behaviour facilitated by structural and psychological conditions. As the thesis argues, particularly in Chapter 5 and 6, such conditions are not incidental but designed and should therefore be subject to regulatory scrutiny through a statutory duty of care model.

Wachs and Wright define online hate speech as including "denigration, harassment, exclusion, and advocacy of violence against specific groups of people on the basis of assigned or selected

[67] Lapidot-Lefler and Barak, above n 37, at 435.
[68] Lapidot-Lefler and Barak, above n 37, at 435.

characteristics" via digital platforms.[69] Each of these terms reflects a distinct mode of harm within the broader ecosystem of online hostility. Denigration involves belittling or demeaning speech aimed at undermining the dignity of targeted groups. While harassment was previously discussed, Wachs and Wright position it within a broader continuum, highlighting its role as repeated, hostile conduct that contributes to the silencing of vulnerable voices. Exclusion captures the structural marginalisation of certain voices from online spaces, while advocacy of violence signals the most extreme form, inciting physical harm.[70] This comprehensive definition is persuasive because it recognises that online hate is not limited to isolated acts of aggression; it often operates through a continuum of speech and silence, where systemic exclusion and symbolic violence are just as impactful as overt threats.

Importantly, their study draws attention to the role of bystanders, who are users who witness online hate but do not intervene. While these users may not directly participate in hate speech, their inaction reinforces its legitimacy and reach. The authors argue that the failure to report, respond to, or condemn hateful content contributes to its diffusion, particularly in algorithm-driven environments where engagement often boosts visibility. [71] This observation is especially relevant to this thesis's focus on platform responsibility. It highlights how platforms design choices can influence user behaviour and collective responses to hate. This reinforces the argument developed in Chapter 5 and 6 that a regulatory response must include duties imposed on platforms to interrupt or intercept, rather than passively accommodate the spread of online hate.

[69] Sebastian Wachs and Michelle F. Wright "Associations between Bystanders and Perpetrators of Online Hate: The Moderating Role of Toxic Online Disinhibition" (2018) 15 International Journal of Environmental Research and Public Health at 1.
[70] Wachs and Wright, above n 69.
[71] Wachs and Wright, above n 69.

Crucially, Wachs and Wright apply Social Learning Theory, later developed into Social Cognitive Theory by Albert Bandura, to explain this process. [72] The theory puts forward that people learn behaviours by observing others, especially when those behaviours appear to be rewarded or go unpunished. This kind of learning occurs through processes of imitation and modelling. Within online spaces, this learning extends to digital cues: if users observe others engaging in hate speech without consequence, or even receiving likes, shares, or attention, they are more likely to adopt similar behaviours.[73] In this way, bystanders are not neutral. Their silence or lack of response can help support these actions, shaping the norms of the online space.

Wachs and Wright's insights are particularly useful for this thesis, as they highlight how hate speech is not solely the result of individual prejudice, but also the product of social conditioning and platform dynamics. Their application of social learning theory in this context, as it complements the Online Disinhibition Effect by adding a collective dimension, indicate that harm spreads not just because users are uninhibited, but because the online environment rewards and replicates antisocial behaviour. This dual lens helps explain both the psychological trigger for toxic behaviour and the broader cultural conditions that allow it to persist. The disinhibited state enables users to act with fewer internal restraints, while the social environment, shaped by algorithms and peer responses, reinforces and normalises such actions.

To the best of my knowledge, this thesis is among the first to connect Online Disinhibition Effect and Social Learning Theory specifically in the context of online hate speech regulation. This theoretical integration, rarely discussed in current literature, forms one of the key

[72] Tim Boone and others "Social Learning Theory Albert Bandura Englewood Cliffs, N.J.: Prentice-Hall, 1977. 247 pp. paperbound Group & Organisation Studies, at 384.
[73] Boone, above n 72, at 384.

contributions of this thesis. It also underscores the need for a regulatory response that addresses not only individual conduct but also systemic incentives, another reason this thesis advocates for a statutory duty of care framework.[74] As discussed earlier, toxic disinhibition operates alongside social learning mechanisms and platform architecture to normalise harmful behaviour online.

To better understand the mechanisms underpinning online hate speech, it is important to engage with foundational theories of deviant behaviour. This is because much of the conduct discussed, whether involving hate speech, trolling or harassment, operates at the boundary of legality and social acceptability. A foundational theory for understanding deviant behaviour is Edwin Sutherland's Differential Association Theory (DAT), developed in the 1920s as part of his broader criminological work.[75] DAT proposes that criminal or deviant behaviour is learned through social interaction, particularly within close peer groups. Individuals adopt the values, techniques, and motivations for deviance from their immediate associations, especially when those interactions favour rule-breaking over rule-following. Crucially, this theory rejects biological or moralist explanations for crime; instead, it frames deviance as socially acquired through communication.[76] Initially developed in the context of youth delinquency and group deviance, DAT remains relevant for understanding how social environments shape individual behaviour, though its applicability to digitally mediated harm is more limited.

Sutherland's model was developed in a pre-digital, proximity-based context, and its relevance to online harm is limited by its emphasis on close, face-to-face associations. In contrast, Albert

[74] To the best of my knowledge, this thesis is among the first to connect Online Disinhibition Effect and Social Learning Theory in the context of online hate speech regulation. This combined approach offers a new way to understand how harmful behaviour spreads and is reinforced through both user psychology and platform design.
[75] Karl-Dieter Opp *Edwin H. Sutherland's Differential Association Theory* (1st ed., Routledge, 2020) at 138.
[76] Cynthia Vinney "Sutherland's Differential Association Theory Explained" (2019) <https://www.thoughtco.com/differential-association-theory-4689191>.

Bandura's Social Learning Theory (SLT), later developed into Social Cognitive Theory, extends the learning process to include symbolic modelling: the idea that individuals also learn behaviours through media exposure, imitation, and the observation of rewards and punishments in others.[77] This is especially applicable to online settings, where behaviour is shaped not only by direct interactions but also by networked observation; where users witness how others behave and how platforms and audiences respond.

SLT has been widely applied to digital environments to explain phenomena such as cyberbullying, hate speech, and aggressive online commentary. As Lowry et al. argue, online platforms facilitate social learning through visibility, reaction metrics (likes, shares), and the perception of peer approval.[78] When users see others engaging in harmful speech without consequence, or even gaining attention, they may interpret these signals as social validation, increasing the likelihood of imitation. In this way, the architecture of social media platforms contributes to a behavioural feedback loop. Importantly, the meaning of a like or share in this context goes beyond simple engagement. A like may serve as a digital endorsement, while a share can function as amplification, whether the intent is to support, ridicule, or simply spread content. These signals are not neutral; they shape how content is interpreted and valued within a digital community. For instance, when hateful posts receive high engagement, this can contribute to the normalisation of such views and create a distorted sense of consensus. While not all engagement implies agreement, the lack of disapproval combined with platform algorithms that reward interaction may result in the increased visibility of harmful content.

[77] Boone, above n 72, at 384.
[78] Paul Benjamin Lowry and others "Why Do Adults Engage in Cyberbullying on Social Media? An Integration of Online Disinhibition and Deindividuation Effects with the Social Structure and Social Learning Model" 2016 27(4) Information Systems Research.

As Rösner and Krämer demonstrate, the influence of group norms and anonymity is magnified when such cues are present. These metrics become part of the digital environment that Bandura describes; one in which is observed behaviour, especially when rewarded, becomes a model for others.[79] This insight reinforces the need to consider platform design in any regulatory responses, a key focus in Chapters 5 and 6 of this thesis.

Together, these theories provide a layered understanding of how harmful behaviours, including hate speech, are sustained and reproduced online. While Differential Association Theory helps explain how prejudiced beliefs may take root within group settings, Bandura's Social Learning Theory (SLT) is more directly applicable to digital spaces. SLT captures how users model behaviour observed online, especially when such actions are visibly rewarded or go unpunished. The integration of both theories suggests that online hate speech is both socially acquired and algorithmically reinforced in that it is learned through group dynamics and sustained by digital validation mechanisms. This theoretical synthesis strengthens the case for a regulatory model that addresses both individual conduct and platform-level incentives.

While Sutherland's Differential Association Theory offers valuable insights into how deviant norms develop within close social groups, it is less equipped to explain behavioural learning in the dispersed, impersonal and media-rich environment of the internet. This is where Bandura's Social Learning Theory (SLT) becomes particularly relevant. SLT builds on the idea of learned behaviour but extends it to include observational learning through indirect exposure, such as watching others online. In digital spaces, users may imitate influencers, viral content creators or even anonymous users, especially when those behaviours are socially validated

[79] Leonie Rösner and Nicole C. Krämer "Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments" 2016 2 Social media + society.

through likes, shares or follower counts. Unlike DAT, SLT recognises the role of symbolic modelling and media cues in shaping behaviour. It is therefore, better suited to explain how online hate can proliferate without close personal contact.

This thesis draws primarily on Social Learning Theory, in combination with the Online Disinhibition Effect, because both theories addressed the psychological and structural factors that sustain hate speech online. SLT helps explain how users imitate harmful behaviours that are visible, reinforced, and unpunished, while ODE highlights the ways in which digital environments lower social inhibitions and accountability. Together, these theoretical lenses provide a structured framework for analysing the adequacy of New Zealand's legal response to online hate speech. This integrated, interdisciplinary framework is a key contribution of the thesis, offering a more comprehensive foundation than doctrinal analysis alone. It also supports the central argument: that hate speech is best understood not merely as individual malice or prejudice, but as harm shaped and amplified by the design and operation of online platforms.[80] While these theories are central to the thesis's behavioural analysis, they are not the only contribution. Later chapters also incorporate legal theory, normative debates, and comparative analysis to develop a more comprehensive case for regulatory reform, particularly through the statutory duty of care framework developed further in Chapters 6.

Building on the theoretical insights of Social Learning Theory and the Online Disinhibition Effect, it becomes clear that online hate speech cannot be understood solely as an individual psychological reaction. Instead, it should be viewed within a broader socio-technical context that includes the interactions between users, technological affordances, and platform

[80] This theoretical framework will be expanded in subsequent chapters, including the development of a statutory duty of care model and its application to platform governance and legislative reform.

architectures. Theories of digital behaviour must therefore be complemented by frameworks that consider the structural and relational aspects of online harm. This is where the work of Tong becomes particularly useful.

Recent interdisciplinary research has extended behavioural accounts of online harm by incorporating relational and structural dimensions of hate speech. Tong offers a typology that classifies online hate into three categories: one-to-one, many-to-one, and intergroup hate.[81] These categories reflect distinct but often overlapping patterns of online abuse. One-to-one hate typically involves direct, targeted hostility between individuals, often exacerbated by personal grievance or perceived anonymity. Many-to-one hate refers to pile-ons, where groups target a single individual, often through coordinated or viral abuse. Intergroup hate operates at a broader level, involving ideologically driven hostility between identity-based communities (e.g. race, religion, gender), and is frequently sustained through memes, hashtags, and community norms.

In the digital sphere, memes and hashtags often serve as vessels for hate speech. Memes can embed stereotypes, symbols or coded languages that dehumanises or demean targeted groups (often under the guise of satire or irony) thereby making hateful content easier to share and harder to regulate. Whereas hashtags, organise content into searchable streams, reinforcing narrative cohesion and amplifying visibility. When hateful memes or hashtags receive engagement (such as likes, shares or trending) platform algorithms interpret this as a social validation, promoting further circulation. These mechanisms contribute to the creation of group-specific norms which are the implicit rules of acceptable expression shaped by repeated exposure, peer reinforcement and platform affordance.

[81] Tong, above n 53, at 30-32.

Tong identifies key platform features that facilitate these forms of harm, including pseudonymity, persistent content, algorithmic visibility, and the lack of effective moderation.[82] These affordances do not merely enable hate, but actively shape how it spreads, enhances its visibility and its perceived legitimacy. For instance, algorithmic systems that prioritise engagement frequently elevate emotionally charged or divisive content, which can amplify hate speech. This aligns with SLT, which argues that users model behaviours they see being rewarded (in this case, through likes, shares or algorithmic promotion). Pseudonymity and visual anonymity, which reduce social accountability, also echo the dynamics of the Online Disinhibition Effect, where users feel psychologically distanced from the consequence of their speech. Tong argues that when hate speech appears unmoderated or even encouraged through engagement, it becomes socially reinforced, especially in intergroup dynamics where hostility can become a form of performative identity. These insights reinforce the idea that online hate is not only a product of individual behaviour, but of platform environments that shape, reward, and legitimise toxic expression.

These insights underscore the limitations of regulatory models focused solely on individual culpability or intent. Tong's work points towards a systemic understanding of harm, in which online hate is seen as the result of platform design and governance failures, not just personal malice. From a regulatory perspective, this calls for a shift away from reactive, criminal law frameworks and towards preventative, duty-based approaches. A statutory duty of care, as proposed in this thesis, would impose positive obligations on platforms to design against foreseeable harms, ensure effective moderation, and make algorithmic decisions more

[82] Tong, above n 53, at 33-34.

transparent and accountable.[83] This aligns with the theoretical movement towards platform governance, which views social media not just as neutral hosts but as active architects of digital behaviour. These behavioural theories demonstrate that online harm is not simply the product of individual malice but emerges from psychological mechanisms and social dynamics shaped by platform design. The next section builds on this insight by turning to regulatory theories, which examine how law, norms, markets, and architecture can address the structural conditions that enable harmful speech.

2.3 Regulatory Theories of Online Harm

This section introduces the regulatory and comparative frameworks that guide the legal analysis in this thesis. These frameworks help evaluate how online harm is shaped not only by legal instruments but by broader systemic, technological and cultural forces. By combining regulatory theory with comparative law, the thesis aims to identify both conceptual limits of existing legal responses and the practical models that could inform reform.

2.3.1 Regulatory Modalities and Lessig's Framework

Building on the psychological and behavioural insights discussed earlier, this section takes a more analytical turn by exploring how legal frameworks intersect with the design and governance of digital platforms. In doing so, I aim to lay the groundwork for evaluating regulatory models that extend beyond traditional legal tools. Lawrence Lessig's Regulatory Theory proposes four distinct forces that influence online behaviour: law, market, social norms, and architecture. These categories, while seemingly abstract, provide a useful way to diagnose

[83] Refer to Chapter 5.

the root causes and potential solutions to online harms.[84] This framework is particularly relevant to social media, where harmful speech is not only governed by formal legal provisions but also shaped by platform design, monetisation structures, and community expectations. Lessig's approach enables a more holistic assessment of online governance, especially in identifying the limitations of relying solely on criminal sanctions.

This section builds on the psychological and behavioural perspectives discussed earlier by turning to the question of regulation in digital spaces. As this thesis argues, legal responses to online hate speech must take into account not only the actions of individuals but also the systems within which those actions occur. Lawrence Lessig's regulatory model provides a conceptual foundation for this. He identifies four key "modalities" of regulation that shape human behaviour in online environments: law, market, norms, and architecture.[85] Each modality operates as a form of constraint, influencing how individuals and platforms behave in both explicit and implicit ways.

Lessig's framework is particularly helpful for analysing social media, where harmful speech is not regulated by law alone. Instead, a combination of technical design (architecture), economic incentives (market), and informal community expectations (norms) also govern what is possible or acceptable online. As Lessig notes, "each of these four constraints can act as a regulator, and the choice between them is not neutral"[86]. This means that the way platforms are built, how they generate revenue, and what behaviours they reward all have regulatory effects, even if they do not come from formal law.

[84] Lawrence Lessig "The New Chicago School" (1998) 27 The Journal of Legal Studies 661 at 664.
[85] Lessig, above n 84, at 664.
[86] Lessig, above n 84, at 664.

For example, a platform that allows users to remain anonymous may reduce accountability and facilitate hostile speech, while one that monetises engagement may reward inflammatory content because it draws attention and generates clicks. These choices reflect the market and architecture modalities. At the same time, social norms, such as what is considered acceptable speech within a community, also shape behaviour, particularly in peer-driven environments like Reddit or Twitter (now X). Legal rules, while still relevant, often operate only after harm has occurred. In contrast, architecture and market incentives shape the conditions under which harm can spread.

Lessig's model therefore supports the central claim of this thesis: that a singular reliance on criminal law is insufficient. Effective regulation must be proactive and must operate across multiple domains. It must consider not just what the law prohibits, but also how platform design can prevent harm, how economic incentives may promote toxicity, and how user norms can be guided through structural or regulatory nudges.

In applying Lessig's theory, this thesis argues for a statutory duty of care that operates not merely as a legal prohibition but as a structural intervention. That is, the duty should compel platforms to design their systems (vis a vis algorithm, moderation tools, and engagement incentives) in ways that anticipate and mitigate harm. This use of legal tools to influence architecture and market design aligns with Lessig's vision of "regulation by code," where the structure of the digital environment becomes a key site of policy enforcement.

2.3.2 Murray's Symbiotic Regulation

To develop this further, Andrew Murray introduces the idea of a "dynamic regulatory model", which refers to an adaptable and responsive form of regulation that evolves alongside changes

in technology and online behaviour.[87] Rather than relying on static legal tools, this model incorporates feedback mechanisms and engages with multiple regulatory actors, including platforms and users. Murray refers to this as a "symbiotic regulatory model," emphasising the mutual adaptation between legal frameworks and the design architecture of digital environments.[88]

Murray's theory responds to the reality that law, when rigid, often lags behind technological innovation. He argues that regulatory success in online environments requires systems that can learn from, and respond to, platform-level developments and user behaviours.[89] This adaptability, grounded in systems theory and cybernetic feedback loops, presents regulation as a process of co-evolution, rather than a top-down imposition.

This emphasis on adaptability aligns closely with the realities of platform governance, where regulation must be able to respond to evolving technologies and user behaviours. Symbiotic regulation aligns with the existing infrastructure of the internet (such as platform code and design features) rather than imposing incompatible external mandates. [90] In this context, success is not measured solely by punitive outcomes, but by the ability of regulatory systems to reduce systemic harm, foster user trust, and create safer online environments through structural accountability.

[87] Andrew Murray *The Regulation of Cyberspace* (Routledge-Cavendish, Abingdon, 2007) at 240.
[88] Paul De Hert and Eugenio Mantovani, "Review of The Regulation of Cyberspace" (2008) 2(1) Studies in Ethics, Law and Technology at 4.
[89] De Hert and Mantovani, above n 88, at 4.
[90] Asma Vranaki, 'Review of The Regulation of Cyberspace: Control in the Online Environment by Andrew D Murray' (2008) 1 Journal of Information, Law & Technology. < https://link.gale.com/apps/doc/A193140958/AONE?u=waikato&sid=bookmark-AONE&xid=77d10612>

However, Murray's approach is not without challenges. A key difficulty lies in coordinating diverse actors with varying interests and capacities.[91] Moreover, symbiotic models rely on a level of institutional trust and transparency that may not always be present in relationships between regulators and powerful technology companies. Despite these concerns, the model remains valuable for its emphasis on feedback, responsiveness, and systemic design.

These insights are central to this thesis's core inquiry: assessing whether New Zealand's current legal framework is capable of addressing the complexity of online hate speech, and whether more dynamic and symbiotic forms of regulation, such as those emerging internationally, offer models for effective reform.

2.3.3 Reidenberg's Lex Informatica and Cohen's Critique

Drawing on Reidenberg's Lex Informatica, this thesis identifies two regulatory models of platform governance: (i) contractual agreements between Internet Service Providers (ISPs) and users, which embed behavioural norms directly into service terms, and (ii) network architecture, where technological design, "code", operates as a regulatory mechanism in its own right.[92] Reidenberg describes this latter form as "a new way of looking at control and regulation in the online environment", whereby the system's architecture effectively enforces policy without the need for traditional legal rules.[93] This "Lex Informatica" model places considerable regulatory power in the hands of developers and platform operators, raising critical concerns about democratic accountability, transparency, and user rights. In the context of social media, this thesis applies Reidenberg's framework to evaluate whether platforms'

[91] Vranaki, above n 90.
[92] Joel R. Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules through Technology*, 76 Tex. L. Rev. 553 (1997-1998) 28 at 555-557.
[93] Reidenberg, above n 92, at 582.

Community Guidelines and design features function as effective regulatory tools for managing online hate speech.

In contrast, Murray advances a dynamic, symbiotic model of regulation that critiques the rigidity of traditional legal frameworks. Murray's approach draws on systems theory and autopoiesis to argue that regulation must co-evolve with the digital environment.[94] Rather than relying on fixed rules, Murray envisions a responsive regulatory system built on continuous feedback loops between regulators, platforms, and users. This "symbiotic" relationship ensures that legal and architectural controls are aligned, increasing the likelihood of effective compliance and systemic resilience.[95] Although more adaptive, this model also presents challenges, including coordination among diverse actors and the difficulty of maintaining consistent regulatory standards.

While Reidenberg's framework offers powerful tools for analysing platform control, this raises urgent questions about the democratic implications of architectural regulation. While Reidenberg's Lex Informatica model provides a compelling account of how online behaviour can be governed through code and contracts, it is not without critique. Julie E. Cohen warns that regulatory reliance on network architecture risks entrenching opaque systems of control that prioritise market interests over public accountability.[96] Her critique is especially relevant for this thesis, which investigates how invisible technical decisions shape user experience and, ultimately, free expression.

[94] De Hert and Mantovani, above n 88, at 5.
[95] De Hert and Mantovani, above n 88, at 7.
[96] Reidenberg, above n 92, at 555.

Cohen argues that when technological design acts as law, it can hide political choices under a technical neutrality façade. This situation presents a real danger; it's eroding democratic oversight and regulatory legitimacy. And yet, in this ostensibly neutral space, so much of the regulation that's happening is increasingly being done via architecture itself.[97] Nonetheless, architectural regulation is increasingly dominant in social media, where Community Guidelines, content moderation tools, and algorithmic filtering form an implicit regulatory regime. This critique highlights the need for transparency and normative scrutiny when considering architecture as a primary mode of governance, particularly in the regulation of online hate speech. These theories demonstrate that the architecture of online platforms is itself a form of regulation that interacts with law and market forces.[98] While regulatory theories explain how different modalities can constrain online harm, they cannot alone resolve the normative question of when and why speech should be restricted. For this, it is necessary to engage with legal and philosophical debates about free expression and its limits. Section 2.4 therefore considers liberal, civic republican, and dignity-based approaches to free speech and hate speech.

2.4 Normative and Legal Theories of Free Speech and Hate Speech

The prevalence of hate speech, both offline and online, is a growing concern globally, and its harmful effects on individuals and communities cannot be overstated. Hate speech is broadly characterised by expressions that are discriminatory, abusive, or threatening and that target individuals or groups based on attributes such as race, religion, gender, or sexual orientation.

[97] Julie E Cohen *Imagining the Networked Society*, in *Configuring the Networked* (Yale University Press 2012) at 5.
[98] Refer to Chapter 5.

This thesis adopts a definition informed by Jeremy Waldron's framework, which highlights the harm hate speech inflicts on both personal dignity and the societal good of assurance.

This aligns with J.S. Mill's foundational defence of free speech in On Liberty, which argues that liberty, including expression, may only be curtailed to prevent harm to others. As Mill writes, "the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others."[99] While Mill was deeply sceptical of censorship, his "harm principle" provides a philosophical basis for restricting speech that crosses into the territory of social or psychological injury.[100] This distinction becomes critical in the online context, where algorithmic amplification can cause disproportionate harm.

Mill's approach to liberty is grounded in individual autonomy but it is not a free-for-all. While he feared the dangers of state censorship, he was equally concerned about the "tyranny of the majority," where dominant social forces could silence dissenting or minority views.[101] In the online context, this insight remains highly relevant: today's digital platforms often amplify the loudest or most sensational voices, while marginalised users may be drowned out, harassed, or excluded altogether. Mill's warning implies that harm is not always physical or immediate, instead, it can also arise from patterns of exclusion, intimidation, or social degradation.[102] Therefore, applying the harm principle to online hate speech requires an expanded understanding of what "harm" can mean in digitally mediated societies. This thus provides a normative justification for considering regulation where such speech prevents equal

[99] John Stuart Mill *On Liberty*, edited by David Bromwich, and George Kateb (Yale University Press, 2003) at 80.
[100] Mill, above n 99, at 80.
[101] Mill, above n 99, at 75-76.
[102] Mill, above n 99, at 80.

participation in public discourse. Defining what is considered "speech" is not always simple. In classical free speech law, the focus is mainly on spoken or written words. In the digital world, however, expression often takes other forms. Memes, cartoons, images, emojis, and even content produced by algorithms can communicate strong ideas. These kinds of expression raise questions for traditional theories, because they move across the line between text, art, and symbolic communication.

Memes, for example, may appear humorous or light, but they can also spread racist stereotypes, promote false information, or encourage hostility. Cartoons too have a long history as political commentary, but they may also be used to send harmful or exclusionary messages.[103] In online spaces these forms can travel quickly, sometimes with more influence than words. This shows that "speech" online cannot be limited only to text.

Yet, later theorists have questioned whether this individual focus is sufficient for democratic participation. For the purposes of this thesis, the term "speech" will be used in a wide sense. It includes not only spoken and written language, but also symbolic, visual, and digital forms of communication that are part of public debate. This approach follows the spirit of international human rights instruments, which protect "expression" rather than words alone. It also makes sure that the study of online hate speech reflects the reality of how harmful content appears and spreads in digital platforms.

Civic republican thinkers such as Alexander Meiklejohn shift the emphasis from individual liberty to collective self-government. Meiklejohn, advancing a civic republican approach,

[103] As later discussed in Chapter 5.5 the Böhmermann case illustrates how satirical expression (in this case, a poem) tests the boundaries between protected expression and legal restrictions.

views freedom of speech not as an individual right alone, but as a structural necessity for democratic self-governance. [104] From this perspective, the protection of public discourse, including its limits, must be evaluated in terms of its contribution to collective decision-making and civic equality.

In online environments, this can take the form of coordinated harassment or algorithmic patterns that amplify harmful content while marginalising vulnerable users. The harm principle (when applied today) must account for this new reality where speech can cause real psychological, reputational, and civic damage not through direct incitement but through exclusion, repetition, and reach.

Waldron develops this further by arguing that hate speech uniquely undermines dignity and public assurance, values essential for equal participation. Waldron's concept of "the public good of assurance" is particularly relevant here.[105] It is not only about protecting individuals from emotional upset but about ensuring that society functions with a shared sense of dignity and inclusion. When speech systematically targets particular groups, denying them equal respect, it erodes the trust that underpins democratic discourse.[106] This is not merely theoretical. Studies show that targeted online hate can discourage civic participation, silence marginalised voices, and contribute to broader social polarisation.

[104] Alexander Meiklejohn, *Free Speech and Its Relation to Self-Government* (Harper & Brothers, 1948) at 26.
[105] Waldron, above n 3, at 34 and 108.
[106] Waldron, above n 3, at 109.

Meiklejohn's work reinforces this. He argues that free speech must be protected because it is necessary for self-government.[107] But for public deliberation to be truly democratic, all voices must be able to participate. If hate speech drives some individuals out of the conversation, the marketplace of ideas fails, not because speech is limited, but because it becomes structurally unequal.

Therefore, this thesis takes the view that freedom of expression must be balanced against the right to dignity and inclusion, particularly in the digital environment. This does not mean that any offensive or controversial speech should be prohibited. Instead, it calls for a regulatory framework that recognises when speech functions not as participation, but as exclusion. This perspective guides the analysis in later chapters, including the adequacy of current New Zealand law (Chapter 4), the comparative and transnational responses (Chapter 5), and the responsibilities of platforms under proposed regulatory models (Chapter 6).

Despite its foundational status, the right to free speech is not absolute. Most legal systems impose restrictions to protect other important interests, including public order, national security, and the rights of others. International law reflects this balance. Article 19 of the International Covenant on Civil and Political Rights (ICCPR) affirms the right to freedom of expression, but allows for limitations necessary for the respect of others' rights or for the protection of national security, public order, public health and morals.[108]

[107] Meiklejohn, above n 104, at 26.
[108] International Covenant on Civil and Political Rights "General comment No. 34, Article 19: Freedoms of opinion and expression" CCPR/C/GC/34, 12 September 2011.

In New Zealand, section 14 of the New Zealand Bill of Rights Act 1990 (NZBORA)[109] protects the right to freedom of expression, while section 5 permits reasonable limits that can be demonstrably justified in a free and democratic society.[110] Courts must therefore engage in a process of proportionality, assessing whether limits on speech are necessary and no more restrictive than required. This will be explored in CHAPTER 4.

It is within this normative and legal framework that the regulation of hate speech must be understood. While offensive speech may be protected, hate speech (speech that is discriminatory, abusive, or threatening and targets groups based on race, religion, ethnicity, gender, or other protected characteristics) raises distinct challenges. As Jeremy Waldron argues, hate speech undermines the public good of assurance, the societal guarantee that all individuals enjoy equal standing and respect.[111] He distinguishes between individual emotional harm and the broader societal damage caused by undermining the dignity of entire groups.[112] Waldron further contends that the harm of hate speech lies not merely in its potential to incite violence but in its constitutive function: it communicates that certain individuals or communities do not belong, eroding the foundations of social inclusion and equality.[113] This thesis adopts Waldron's approach, arguing that hate speech is a unique form of harm that warrants targeted regulatory responses.

2.4.1 Conceptual Foundations of Free Speech and the Threshold for Hate Speech

[109] New Zealand Bill of Rights Act 1990, s 14.
[110] New Zealand Bill of Rights Act 1990, s 5.
[111] Waldron, above n 3, at 109.
[112] Waldron, above n 3, at 5.
[113] Waldron, above n 3, at 106.

Before exploring the limits of free speech, it is necessary to understand what freedom of speech is for, why it is valued and how it contributes to a democratic society. Different theories of free expression offer distinct justifications for its protection, ranging from the promotion of individual autonomy to the maintenance of an informed public. These foundational ideas shape how we assess speech and when, if ever, it may be justifiably restricted. The following section outlines key conceptual frameworks that help define the purpose and function of free speech, setting the stage for a more nuanced discussion of its legal boundaries and its intersection with hate speech.

2.4.2 What is Free Speech for?

One of the most influential liberal theories of freedom of expression is John Stuart Mill's argument in On Liberty. Mill explains that individual liberty, including freedom of speech, may only be limited to prevent harm to others. This is known as the "harm principle".[114] It sets a high bar: the government or others should not restrict a person's speech just because it is unpopular or offensive. Instead, there must be clear and serious harm.

Although Mill is often used to defend strong free speech rights, his theory also supports regulation in some cases. Mill was worried not only about government censorship but also about the "tyranny of the majority" where popular opinion can silence minority or unpopular views.[115] Today, this is relevant online, where dominant voices or aggressive groups can exclude or intimidate others. For example, harmful speech can be amplified through algorithms, making it harder for targeted groups to feel safe or included.

[114] Mill, above n 99, at 80
[115] Mill, above n 99, at 80.

Mill's harm principle is not only about physical harm. It can also include psychological, reputational, or civic harm. These are the kinds of harm often seen with online hate speech. In digital spaces, harmful speech can be repeated and shared widely. It can make certain people or communities feel they do not belong. This causes exclusion and can stop people from joining public conversations.

Therefore, using Mill's theory today means thinking about how online speech can cause long-term harm. It also means thinking about who is being silenced or driven away from public spaces. This helps support a fairer and more democratic digital environment, where everyone can participate.

This approach connects with later sections of this thesis, including Waldron's argument that hate speech damages dignity and inclusion[116], and Meiklejohn's view that speech should support democratic self-government.[117] Together, they help to demonstrate why freedom of expression must sometimes be limited, not to control ideas, but to protect equal participation and prevent harm.

Most legal systems do not treat freedom of expression as absolute. As noted earlier, Mill's harm principle provides a useful threshold; however, contemporary legal frameworks often refine this through proportionality analysis. For instance, Article 19(3) of the International Covenant on Civil and Political Rights (ICCPR) recognises freedom of expression as a fundamental right but permits restrictions where they are necessary and proportionate to protect the rights or reputations of others, national security, public order, public health and morals.[118]

[116] Waldron, above n 3, at 5 and 108.
[117] Meiklejohn, above n 104, at 26.
[118] International Covenant on Civil and Political Rights "General comment No. 34, Article 19: Freedoms of opinion and expression" CCPR/C/GC/34, 12 September 2011.

This balance has been reinforced in the work of UN Special Rapporteurs on the promotion and protection of the right to freedom of opinion and expression which stress that limitations on expression must be lawful, necessary, and proportionate to a legitimate aim. In the context of hate speech, such regulation must be carefully designed to prevent harm while preserving space for open debate.[119] These standards, drawn from Article 19(3) of the ICCPR, are further examined in Chapters 4 and 5, where they help evaluate New Zealand's legal framework and the responsibilities of online platforms.

2.4.3 Limits of Free Speech in Law

It is also important to examine how different legal traditions interpret the boundaries of free speech. In the United States, the First Amendment provides strong protections for expression, even in cases involving speech that may be considered hateful or offensive. The U.S. Supreme Court has generally adopted an "absolutist" approach, where only a very narrow set of exceptions, such as incitement to violence, defamation, or obscenity, justify government restriction. For example, in *R.A.V. v City of St Paul*, the Court invalidated a local ordinance that prohibited hate symbols, holding that even racially charged expression could not be banned based on viewpoint.[120] In *Snyder v Phelps*, the Court upheld the right of a group to protest with offensive signs near a soldier's funeral, emphasising that public speech on matters of concern cannot be punished simply because it causes distress.[121] This absolutist stance illustrates a sharp contrast with the balancing approach adopted under international human rights law and by New Zealand courts. This model contrasts with the approach taken under international human rights

[119] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UN Doc A/74/486 (9 October 2019).
[120] *R.A.V. v City of St Paul* 505 US 377 (1992).
[121] *Snyder v Phelps* 562 US 443 (2011).

law and in New Zealand, where speech can be limited more readily when it conflicts with other rights or social values.

New Zealand law reflects this balancing model. Section 14 of the New Zealand Bill of Rights Act 1990 (NZBORA) affirms the right to freedom of expression, while section 5 allows for reasonable limits that are demonstrably justified in a free and democratic society. Courts have developed a structured approach to proportionality, assessing whether such limits pursue a sufficiently important aim and impair the right no more than necessary.

This domestic legal framework aligns with Article 19 of the International Covenant on Civil and Political Rights (ICCPR), which guarantees freedom of expression while allowing restrictions where they are lawful, necessary, and proportionate to protect the rights or reputations of others, national security, public order, public health and morals. These international standards provide an important reference point for evaluating the legitimacy of hate speech regulation.

2.4.4 Proportionality and the Justification for Limitations

Stephen Breyer's theory of proportionality provides a more judicially grounded account, showing how courts can balance free speech with democratic fairness. In addition to Mill's liberal account of harm, a more democratic rationale for limiting speech is offered by legal jurist, Stephen Breyer. This section draws on the work of Breyer to explore how proportionality supports a more nuanced and democratic approach to limiting speech. Later chapters will

consider how these balancing approaches have been taken up in New Zealand jurisprudence, particularly through the s 5 proportionality test under the New Zealand Bill of Rights Act 1990.

A complementary theoretical perspective can be found in Stephen Breyer's concept of "active liberty", which further supports a democratic rationale for certain limits on expression. Breyer's idea of active liberty adds an important layer to the discussion about the purpose of free speech in a democratic society. In Active Liberty: Interpreting Our Democratic Constitution, Breyer explains that freedom of speech should not be understood only as protection from government interference, but also as a democratic tool that helps people participate in public decision-making.[122] From this perspective, speech is valuable not only because it allows individuals to express themselves but also because it supports inclusive public discussion and civic equality.[123]

Breyer's theory is especially useful for thinking about hate speech in online environments. He argues that courts and lawmakers should consider how speech affects democratic participation.[124] Some limits on speech may be justified, not because they suppress expression, but because they protect the democratic process.[125] For example, laws that prevent users from being harassed or excluded from online spaces may help ensure that everyone has a fair opportunity to participate in public life. In this sense, Breyer's framework suggests that certain restrictions on speech can support democracy, particularly when they protect people who may otherwise be silenced.

[122] Stephen Breyer *Active Liberty: Interpreting Our Democratic Constitution* (The Tanner Lectures on Human Values, Harvard University, 2004) at 10.
[123] Breyer, above n 122, at 22.
[124] Breyer, above n 122, at 8 and 27.
[125] Breyer, above n 122, at 10 and 27-28.

Breyer's approach is closely connected to the idea of proportionality. In New Zealand, this principle is central to how courts apply s 5 of the New Zealand Bill of Rights Act 1990. When courts assess whether a limit on free speech is justified, they weigh the importance of the right against the harms the speech may cause. Breyer's concept of active liberty supports this kind of careful balancing; it suggests that the impact on the broader community (including the effects on democratic participation, public trust, and inclusion) should also be considered.[126]

At the same time, Breyer does not suggest that all restrictions are justified. He emphasises that any limit must be reasonable and necessary; he also argues that the courts should pay close attention to the real-world consequences of speech, as opposed to relying solely on abstract legal principles.[127] This is especially important in the United Sates, where courts often take a strong protective approach to free speech. Even within that context, Breyer has argued that restrictions may be appropriate where they serve to protect the fairness of democratic processes; such in the case of *McConnell v Federal Election Commission*.[128] In that case, Breyer's reasoning further illustrates his commitment to balancing free speech with democratic participation.[129] In that case, the United States Supreme Court upheld key provisions of the Bipartisan Campaign Reform Act 2002 which placed limits on 'soft money' contributions to political parties. Breyer supported the majority view that these limits were justified; not because political speech is unimportant, but because unchecked financial influence could distort the political process and reduce public confidence in democratic institutions.[130] For Breyer, the restriction served to protect 'active liberty'; the capacity of citizens to engage

[126] Breyer, above n 122, at 8 and 33.
[127] Breyer, above n 122, at 11-12 and 33.
[128] *McConnell v Federal Election Commission* 540 US 93 (2003), as discussed in Breyer, above n 122, at 32-34.
[129] *McConnell v Federal Election Commission* 540 US 93 (2003).
[130] *McConnell v Federal Election Commission* 540 US 93 (2003), as discussed in Breyer, above n 122, at 32-34; at 137-138.

meaningfully in democratic decision-making.[131] This type of reasoning, which focuses on the real-world consequences of speech for democratic equality, supports regulatory approaches that aim to ensure fairness in online environments too.

Breyer's reasoning highlights that the U.S. Supreme Court applies different tests when assessing limits on free speech, sometimes permitting restrictions where the integrity of democratic processes is at stake. His approach contrasts with the common view that U.S. jurisprudence adopts an absolutist position. Importantly, Breyer's focus on real-world consequences exposes a useful parallel for analysing online environments. Media corporations often invoke the language of free speech to resist regulation; however, their deeper concern lies in preserving engagement and advertising revenue. In other words, their defence of "speech" resembles the role of unchecked financial influence in campaign funding: both risk distorting democratic participation. Linking Breyer's reasoning with behavioural theories of online communication, such as the Online Disinhibition Effect and engagement-driven amplification, allows this thesis to frame platform resistance to regulation not as principled defence of liberty, but as a protection of profit.

Breyer's theory helps move the debate beyond the binary of whether speech should be permitted or prohibited. Instead, it asks a deeper question: does the speech environment support democracy, fairness, and participation? This is especially important in the context of online hate speech, where the potential to undermine civic inclusion and democratic values is significant.

[131] Breyer, above n 122, at 33.

This perspective is also relevant to speech on social media. Today, digital platforms act as modern public forums. They shape who is heard and which messages are amplified. As a result, online hate speech can have serious consequences for democracy. It may discourage some individuals or groups from speaking at all, especially those from marginalised communities. Breyer's work helps explain why regulation in these spaces should aim to create a fair environment where all people can participate in public discussion.[132] This reasoning underpins the later analysis of platform duties in Chapter 6, where the need to maintain equitable digital participation is a central concern.

2.4.5 What qualifies as "Hate Speech"?

International definitions typically describe hate speech as expression that incites discrimination, hostility or violence against individuals or groups based on characteristics such as race, religion, gender or sexual orientation. For example, the United Nations defines hate speech as "any kind of communication in speech, writing, or behaviour that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of their identity".[133] While this provides a general threshold, such definitions often lack the depth needed to guide legal or constitutional responses.

"In every society, there is a group of people who suffer from bigotry and ignorance and in every society, it is easy for the socially disenfranchised to blame another, social, racial or religious group"[134]

---

[132] Breyer, above n 122, at 10 and 34.
[133] United Nations "United Nations Detailed Guidance on Implementation for United Nations Field Presences" (2020) <https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf>), p.2
[134] Andrew Murray *Information technology law: the law and society* (4th ed, Oxford University Press, Oxford 2019) at 128.

A precise, universally recognised definition of hate speech remains elusive in international and domestic legal contexts. While scholars broadly agree that hate speech is characterised by its discriminatory, abusive, and threatening nature (often targeting individuals based on ethnicity, religion, disability or sexual orientation) this conceptual ambiguity complicates regulatory efforts.[135] These theoretical insights raise pertinent questions about how legal systems can define and respond to hate speech in practices. The difficulty of drawing precise boundaries without undermining freedom of expression is explored in later chapters through the lens of New Zealand law.

This section builds on those foundations by drawing on Jeremy Waldron's theory of dignity to examine the deeper societal harms of hate speech. Waldron argues that dignity is not merely about personal esteem but about one's standing as a member of society, entitled to respect and equal treatment.[136] Hate speech undermines this standing, creating an environment of exclusion and distrust.[137] For example, Waldron distinguishes between societal harm, where the collective fabric of equality is damaged, and individual harm, which includes emotional distress or reputational damage.[138] Waldron further categorises the impacts of hate speech as either consequential, resulting in tangible acts like violence, or constitutive, where harm is inherent in the act itself.[139] This theoretical lens provides a foundation for evaluating how legal systems, including New Zealand's approach, approach the regulation of hate speech, a key theme in later chapters, which examine how legal systems, including New Zealand's, respond to these forms of harm.

[135] Robert Mark Simpson "Dignity, Harm, and Hate Speech" (2013) 32 Law and Philosophy 701 at 727.
[136] Waldron, above n 3, at 5.
[137] Waldron, above n 3, at 105.
[138] Waldron, above n 3, at 115.
[139] Waldron, above n 3, at 34 and 108.

Waldron argues that distinguishing hate speech from merely offensive speech is critical in legal and policy frameworks.[140] While both types of speech may evoke emotional responses, hate speech uniquely undermines societal norms of equality and inclusion, thereby harming the public good. Rejecting the idea that legal distinctions in this area must be overly simplistic or devoid of ambiguity, Waldron advocates for a nuanced approach.[141] This nuanced approach necessitates a balance between the constitutional value of freedom of expression and the equally fundamental value of human dignity.[142]

According to Waldron, this balance is not mechanical, but interpretive, requiring lawmakers and courts to weigh the communicative value of expression against the societal harms it may cause.[143] Crucially, not all offensive or controversial speech justifies restriction; the threshold for legal intervention must be tied to whether the speech impairs the assurance of equal social standing. This involves considering several elements: the targeted nature of the speech, its public dissemination, its capacity to incite exclusion or fear, and whether it undermines the normative framework of inclusivity. The moral judgment Waldron speaks of is thus grounded in a legal culture that prioritises democratic legitimacy and social cohesion, without resorting to blanket censorship. This thesis adopts that view, arguing that hate speech regulation should be proportionate, precise, and embedded within a rights-based legal system that protects both liberty and dignity.

[140] Waldron, above n 3, at 115.
[141] Waldron, above n 3, at 115.
[142] Waldron, above n 3, at 5.
[143] Waldron, above n 3, at 115.

Waldron's perspective underscores the necessity of addressing not just the emotional or reputational harm experienced by individuals but the broader societal damage that arises when hate speech erodes trust and inclusivity.[144]

2.4.6 Where to Draw the Line? Threshold and Conceptual Tests

The question of where to draw the boundary between lawful and unlawful speech is one of the most difficult issues in hate speech regulation. It is not enough to rely on the definitions of social media companies, because these are often narrow and lack accountability. A stronger approach is to look at legal principles, human rights standards, and academic theories to see what kind of tests can guide the line. One starting point is Mill's harm principle, which makes a difference between speech that only causes offence and speech that produces real harm. According to Mill, offence is not enough to limit speech. Regulation is justified only when speech causes actual harm to others or to society. Waldron develops this idea further. He argues that hate speech becomes unlawful when it damages human dignity and the sense of public assurance.[145] The harm here is not only about violence but about creating an environment where some groups no longer feel included as equal citizens. For Waldron, this is the threshold because it affects the basic conditions of a democratic society.

Alexander Brown adds another layer by saying that definitions of hate speech must be precise and clear. He warns against broad or vague definitions that risk silencing legitimate debate[146]. Instead, Brown focuses on speech that attacks people for belonging to certain protected groups

[144] Waldron, above n 3, at 115.
[145] Waldron, above n 3, at 115.
[146] Alexander Brown "What is Hate Speech? Part 1: The Myth of Hate" (2017) 36 Law and Philosophy at 428.

and has the effect of exclusion or stigmatisation. This precision makes sure the law responds to real harm without going too far.

International human rights law also gives a proportionality test. In New Zealand, the Bill of Rights Act 1990, section 5, requires that limits on rights must be reasonable, serve a clear aim, and interfere with rights as little as possible. Applied to hate speech, this means that restrictions must be carefully tailored: they cannot be too weak, but they also cannot censor more speech than necessary.

In the online space, researchers point out that the digital environment changes the test. As Gagliardone and others explain, online hate speech spreads faster and more widely because of anonymity and algorithmic amplification.[147] These features mean that thresholds developed for offline speech may not capture the full scale of harm online.

Considering all that together, these tests show that the line should not be at the point of simple offensiveness. The stronger position is that the line must be drawn where speech undermines dignity, equality, and participation in society. This thesis uses these principles as the framework for evaluating how New Zealand's current law responds to online hate speech, and whether reform is needed.

2.4.7 Historical and Sociological Dimensions of Hate Speech

Before turning to the normative justifications for regulating hate speech, it is important to situate these debates within their historical and sociological context. Understanding how hate

[147] Iginio Gagliardone and others *Countering online hate speech* (UNESCO, 2015) at 8.

speech has evolved and operated within different social settings helps clarify why its regulation continues to pose such complex challenges. As later chapters will show, jurisdictions differ in how they apply these principles. While some prioritise robust protections for speech, others recognise a broader role for law in maintaining the dignity and equality of all participants in public discourse. These insights form part of the conceptual foundation for the framework developed in the next section, which assesses when and how speech should be regulated in order to protect both liberty and societal cohesion.

Hate speech, while amplified in the digital age, is not a new phenomenon. Numerous historical and cultural examples illustrate the destructive effects of bigoted speech in the real world. Such behaviour can be traced back to fundamental aspects of human nature, where the propensity to harbour animosity towards anything perceived as unfamiliar, inexplicable, or novel is an innate characteristic.

From ancient civilisations to contemporary societies, hate speech has taken various forms targeting individuals or groups based on diverse characteristics such as ethnicity, religion, gender, or other distinguishing characteristics. It manifests itself in politics, religion, social dynamics, and virtually every other aspect of human interaction. This historical continuity demonstrates the enduring nature of hate speech as a regrettable aspect of human behaviour.[148]

The propensity to despise or fear the unfamiliar is a profoundly rooted characteristic of human psychology.[149] Unfamiliar entities or circumstances may pose a threat to one's health or group

[148] Laurent Pernot *Epideictic Rhetoric in Ancient Greece* (University of Texas Press, Austin, 2015) at 25.
[149] Agnieszka Pluta and others "Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain" (2023) 13 Scientific Reports at 9.

cohesion.[150] In this context, the tendency to mistrust and reject those who are perceived to be distinct may have served as an adaptive mechanism in ancestral environments.

As societies evolved and became more diverse, this innate tendency to detest the "other" has persisted, perpetuating stereotypes, prejudices, and discriminatory attitudes. Cultural and societal norms significantly influence the expression and propagation of hate speech. Through socialisation processes, individuals internalise biased perspectives from their surroundings, embedding prejudiced beliefs and discriminatory behaviours.[151]

Modern advances in communication technology, particularly the emergence of the Internet and social media platforms, have facilitated the spread of hate speech in new ways. Online anonymity often emboldens individuals to express hateful sentiments, contributing to the global proliferation of harmful content.

To develop effective strategies for combating this pervasive issue, it is essential to understand the historical continuity and psychological foundations of hate speech. Research in sociology, psychology, and communication studies helps illuminate the complex interplay between individual bias, social dynamics, and technological design that contribute to the perpetuation of online hate.[152] Identifying the underlying causes and mechanisms is crucial for developing targeted interventions and educational initiatives to promote empathy, tolerance, and inclusivity.[153] These interdisciplinary perspectives are integrated throughout this thesis, particularly in Chapter 2, which introduced the Online Disinhibition Effect and Social Learning

[150] Pluta and others, above n 149, at 10.
[151] Dominic Abrams "Processes of prejudice: Theory, evidence and intervention. Equality and Human Rights Commission Research report" (2010) 56 Center for the study of Group Processes, University of Kent at 41.
[152] Ariadna Matamoros-Fernández and Johan Farkas "Racism, Hate Speech, and Social Media: A Systematic Review and Critique" (2021) 22 Television & New Media at 205.
[153] Matamoros-Fernández and Farkas, above n 152.

Theory as explanatory frameworks for harmful online behaviour. Additionally, Chapter 6 draws on these insights to support the case for a statutory duty of care for social media platforms, arguing that understanding user psychology and platform architecture is essential to designing effective accountability mechanisms.

In democratic societies, where free speech and open debate are cherished, a notable recurring element in public discourse is what can be termed "blame rhetoric." [154] Rooted in the teachings of the Ancient Greek philosopher Aristotle, rhetoric serves a similar function to dialectic thinking.[155] From this emerged the notion of Epideictic Rhetoric, which pertains to speech involving *praise* and *blame*.[156] In special occasions like weddings or eulogies, the use of Epideictic Rhetoric focuses on accentuating positive attributes and virtues. [157] Aristotle identified virtues such as courage, loyalty, magnificence, reason, magnanimity, and self-control as exemplary traits embodied in praise speeches.

Conversely, “blame rhetoric”, the other side of Epideictic Rhetoric, emphasises vices or negative traits.[158] A historical example of this is the public ridicule of Italian Fascist leader Benito Mussolini after his death. Partisans displayed his corpse for public stoning, symbolising an act of blame rhetoric, where his perceived vices were amplified to alienate and condemn. [159] Blame rhetoric underscores how it can divide audiences and exclude certain groups from a sense of community.[160]

[154] Murray, above n 134, at 115.
[155] Aristotle - Translated by W. Rhys Roberts *Rhetoric* (350 B.C.E).
[156] Mary Hedengren "Epideictic Rhetoric" (podcast, 5 November 2015) Mere Rhetoric <https://mererhetoric.libsyn.com/epideictic-rhetoric>.
[157] Hedengren, above n 156.
[158] Hedengren, above n 156.
[159] Didier Musiedlak "The Metamorphoses of Mussolini’s Body" 2018 20 Journal of Genocide Research 236 at 237.
[160] Hedengren, above n 156.

In essence, epideictic rhetoric reflects the values that a group or audience holds dear. While praise rhetoric reinforces shared norms, blame rhetoric exposes violations of those norms and can serve to exclude or stigmatise. This dynamic is particularly relevant in discussions of online hate speech, where digital expression often performs a rhetorical function; affirming in-group ideologies or vilifying perceived outsiders. Within this thesis, epideictic rhetoric provides a useful interpretive lens for understanding the social function of hate speech in online environments: it helps explain how harmful expression both reflects and constructs collective moral boundaries.[161]

These historical lessons resonate with today's challenges, particularly in the digital sphere, where the amplification of hate speech through social media platforms creates new regulatory challenges. Nearly a century later, while the mediums of communication have evolved, the core mechanisms of hateful expression remain unchanged. The advent of personal computers and the internet in the 1980s introduced new avenues for the mass propagation of hate rhetoric. Modern technologies, including social media and digital platforms, have made it easier than ever for harmful messages to reach global audiences, perpetuating the enduring challenge of combating hateful expression in democratic societies. Waldron's analysis resonates with contemporary debates, where online rhetoric often corrodes public assurance and dignity by reinforcing hierarchies of belonging and exclusion.[162]

[161] For instance, the National Socialist German Workers' Party (NSDAP) weaponised anti-Semitic propaganda, framing "Judeo-Bolshevism" as a perceived threat aligned with far-right ideology. Public speeches and radio broadcasts became key tools in spreading their message to the masses, demonstrating how radio and television served as vehicles for mass propaganda.

[162] For example, Destiny Church leader Brian Tamaki has increasingly used social media to disseminate divisive messages targeting LGBTQ+ people, Māori, and religious minorities, under the guise of religious or political commentary. His digital communication strategy deliberately appeals to marginalised frustrations while invoking conspiratorial and exclusionary language. The viral nature of Tamaki's rhetoric also illustrates how platform design and algorithmic amplification can contribute to the spread of socially corrosive content.

2.4.7.1 Modern Challenges

Regulating hate speech in the digital age presents unique challenges stemming from the dynamics of digital communication, where anonymity, virality, and global reach amplify the harm caused by hateful expression. These challenges are particularly acute in New Zealand, where incidents like the Christchurch Mosque shootings have revealed gaps in the legal framework's ability to prevent and address hate speech effectively. As Waldron highlights, hate speech undermines both individual dignity and the public good of assurance, eroding societal trust and inclusivity.[163] Such harms are exacerbated when digital platforms fail to adequately moderate harmful content or when perpetrators exploit online anonymity to evade accountability.

The Internet & Jurisdiction Global Status in 2019, identifies three areas of growing concern in the digital sphere: security, economy and expression. While security issues include cybercrime and data breaches, and economic concerns relate to intellectual property and e-commerce, expression focuses on hate speech, cyberbullying, fake news, and other forms of harmful content. Hate speech, as a subset of harmful expression, contributes significantly to social harm by fostering division, exclusion, and hostility within digital spaces. [164]

From a regulatory theory perspective, this highlights the failures of reactive models. Current enforcement under the Harmful Digital Communications Act 2015 is ill-equipped to address ideologically driven speech that incites cultural division without overt threats. Furthermore, the group's use of online advertising and algorithmic reach illustrates how platform design

[163] Waldron, above n 3, at 85
[164] Dan Svantesson *Internet Jurisdiction Global Status Report 2019: Key Findings* (2019) at 73.

(code) and market forces, per Lessig's model, contribute to the spread of harmful ideas in ways the law has yet to fully address.[165]

In New Zealand, these challenges are illustrated by groups such as the Hobson's Pledge, whose digital campaigns reveal how current legal tools fall short. Examples like Hobson's Pledge should prompt a reconceptualisation of how harm is defined in the digital space. New Zealand's regulatory regime, while rightly protective of political freedom, must account for how repeated and racially charged online messaging can impair democratic participation and social cohesion, especially for Māori communities. This does not require blanket censorship but rather targeted reform, including clearer definitions of harm, obligations on platforms, and mechanisms that address the cumulative effects of exclusionary narratives.

2.4.8 Normative Justifications for Regulating Hate Speech: Dignity, Equality, and Harm

Understanding hate speech as a legal and social problem requires more than behavioural insight or regulatory design; it requires an account of the moral and political values that shape how a democratic society should respond to harm and exclusion. Normative theories help explain why hate speech matters; not simply because it offends, but because it communicates and reinforces social inequality, exclusion, and marginalisation.

This section draws on two key theorists: Jeremy Waldron, who frames hate speech as a threat to human dignity and public assurance, and Sandra Fredman, who provides a legal framework for understanding and responding to structural inequality through the lens of substantive

[165] Robin Barendze "The (In)visibility of Hobson's Pledge: A Struggle for Survival in the Sociopolitical Environment of Aotearoa/New Zealand" (Master of Arts in Sociology, Massey University, New Zealand, 2018) at 87.

equality. Together, their work justifies a shift away from simplistic offence-based approaches and toward regulatory models that seek to uphold equal participation, recognition, and protection in both public and digital spheres.

Jeremy Waldron argues that the central harm of hate speech lies in its denial of dignity and equal membership within the community. Waldron defines dignity as a social status that entitles all individuals to basic respect and protection by public institutions.[166] When hate speech circulates in public environments (whether through posters, graffiti, or social media posts) it signals that certain groups do not belong, and that their equal standing is not socially or institutionally affirmed.[167]

Waldron distinguishes this from mere emotional distress. The problem, he argues, is not just that hate speech causes offence, but that it erodes the public good of assurance; the shared understanding that all members of society are entitled to equal treatment and social respect.[168] In other words, hate speech undermines the basic trust that citizens should be able to place in their communities and its legal order.

This is particularly relevant in digital environments like social media, where the speech is both pervasive and persistent, and where the architecture of platforms often fails to protect marginalised users. In New Zealand, the spread of racist, homophobic, or religious discriminatory speech online has been shown to affect not only the psychological well-being of individuals, but also their willingness to participate in online public discourse.[169]

[166] Waldron, above n 3, at 5.
[167] Waldron, above n 3, at 94-95.
[168] Waldron, above n 3, at 106.
[169] Advice for New Zealanders (12 December 2019) Netsafe <https://www.netsafe.org.nz/2019-online-hate-speech-insights>

Waldron's approach helps to explain why legal regulation may be necessary not to protect people's feelings, but to uphold their status as equal members of society. Waldron's theory thus provides a compelling normative justification for regulatory intervention. It supports the view that law has a role in preserving a public culture in which all people can reasonably expect to be treated with respect. While criminalisation may be one response, Waldron's emphasis on public assurance also opens the door to non-punitive models, such as a duty of care for platforms, that aim to prevent harm by structuring digital environments more responsibly.

While Waldron's theory is a powerful defence of public dignity, it is not without critique. His emphasis on communal assurance risks prioritising social cohesion over individual autonomy, especially in pluralistic societies. This concern becomes more pronounced when viewed alongside Mill's harm principle, which cautions against restricting expression unless it causes clear and demonstrable harm to others, and Breyer's proportionality approach, which requires careful balancing between competing rights. There is therefore a risk that dignity based approaches, if framed too broadly may inadvertently suppress controversial or minority viewpoints in the name of civility. Nevertheless, this thesis adopts Waldron's position as a normative foundation, while remaining cautious about overly broad or vague restrictions that could infringe on legitimate expression.

2.4.8.1 Substantive Equality and Structural Harm

Finally, Sandra Fredman situates hate speech in the broader context of structural inequality, adding a substantive equality lens that complements dignity-based accounts. While Waldron's

theory offers the thesis a principled rationale for limiting harmful speech, Fredman provides a legal framework for understanding how regulation should be designed. Fredman critiques formal equality models that focus narrowly on sameness treatment and instead proposes a multidimensional theory of substantive equality, which takes into account the real-world contexts in which discrimination and exclusion occur.[170]

Fredman's model comprises four dimensions of substantive equality: first, redressing disadvantage, which addresses material and structural forms of exclusion; second, recognising and challenging stigma and stereotyping, which targets cultural and representational harms; third, enhancing participation and voice, which ensures that all groups can contribute meaningfully to public and institutional life; and fourth, facilitating structural change, which focuses on reforming institutions and systems to accommodate difference and dismantle embedded inequalities.[171]

Each of these dimensions helps identify how hate speech contributes to, and reinforces, systemic inequality. First, it disproportionately targets communities that already experience social and economic marginalisation. Second, it entrenches negative stereotypes and reproduces stigmatising narratives. Third, it silences those who are attacked, diminishing their ability to speak, organise, or engage in political life. Fourth, it reflects deeper failures in institutional design, particularly in the case of social media platforms, where algorithmic amplification and uneven moderation often exacerbate online harm.

[170] Fredman, above n 44, at 735.
[171] Fredman, above n 44, at 726-728.

Fredman's model has been widely influential, but it also raises questions about how to translate normative theory into practical lawmaking. As Fredman herself acknowledges, the four-dimensional approach is ambitious and may risk becoming too expansive or conceptually diffuse when applied to real-world regulation.[172] In contexts like digital governance, there may be tension between the goals of structural change and the legal principle of proportionality. However, this thesis contends that Fredman's framework is particularly well-suited to the digital context, precisely because it recognises that equal participation and representation cannot be achieved through narrow legal tools alone. Her work helps justify a model of regulation that is both preventive and systemic, rather than solely punitive.

Fredman's framework is particularly valuable in assessing regulatory responses. Her work suggests that an effective legal approach should not only prohibit certain kinds of expression but should also proactively reshape institutions, including digital platforms, to promote equality and inclusion. In this sense, her theory supports this thesis's proposal of a statutory duty of care, a structural obligation on social media platforms to prevent foreseeable harm, create inclusive user environments, and ensure procedural fairness. Rather than relying solely on criminal penalties, a duty of care approach aligns with Fredman's call for transformative legal design. This is a model that recognises the complexity of inequality and uses regulation to promote meaningful change.

Together, Waldron and Fredman offer a powerful normative foundation for rethinking how the law should respond to online hate speech. Waldron explains the moral harm, namely that hate speech undermines public dignity and corrodes the shared assurance that all are equal members

[172] Sandra Fredman "Human Rights Transformed: Positive Rights and Positive Duties" (2008) 38 Oxford Legal Studies Research Paper 1 at 16-18.

of the community. Fredman provides the legal pathway forward, offering a multidimensional framework for building equality into regulatory design, particularly through institutional duties.

Their theories support the broader contribution of this thesis's central claim: the development of a regulatory framework that goes beyond reactive legal measures to address the systemic conditions enabling online hate speech and harmful content. By virtue of integrating behavioural, regulatory and normative insights, this thesis offers a structured and original way to assess legal responses in Aotearoa New Zealand and elsewhere. This interdisciplinary approach underpins the analysis in the next chapters which examine New Zealand's current legal framework settings and emerging global models of regulations.

These normative frameworks help clarify the values that should underpin any evaluation of legal effectiveness. They support the view that a regulatory framework must not only prevent harm, but must also promote inclusion, protect dignity, and enable equal participation in public life. In doing so, they help guide this thesis's analysis of whether New Zealand's current legal framework meets these standards, and whether a statutory duty of care could provide a more appropriate, proportionate, and forward-looking reform.

2.5 Conclusion and Reflections

This chapter has explained three perspectives that form the base of the thesis. Rather than treating behavioural, normative, and regulatory theories as separate explanatory frameworks, the thesis integrates these perspectives to analyse online hate speech as a socio-technical phenomenon requiring coordinated legal and institutional responses. The first perspective is behavioural. Theories such as the Online Disinhibition Effect, Social Learning Theory, and Differential Association Theory show how online spaces encourage harmful speech. People often feel less accountable when they are anonymous, and behaviour can be repeated and copied in groups. Hate speech online is therefore not only about individual choice but also about wider social influence and the design of platforms.

The second perspective is regulatory. Lessig's theory of regulation through code, Murray's idea of "symbiotic regulation," and Reidenberg's concept of Lex Informatica all demonstrate that technology has its own regulatory force. Platform design, algorithms, and market incentives shape what people see and how they act. This means platforms are not neutral spaces. They regulate behaviour alongside the state and the market. Law must therefore consider not only individual acts of speech but also the structures that make some speech louder and other speech less visible.

The third perspective is normative and legal. Academics such as Mill, Waldron, Meiklejohn, Breyer, and Fredman explain both the importance of free expression and the reasons for limiting it. Mill's harm principle values freedom, but Waldron argues that hate speech attacks dignity and weakens equal membership in society. Meiklejohn and Breyer stress the democratic need to protect meaningful participation, and Fredman links regulation to the value of substantive equality. Taken together, these theories support restrictions on hate speech where it threatens dignity, equality, or democratic life.

The historical and sociological material in Section 2.4.7 shows that hate speech is not a new problem. Past examples of propaganda and vilification reveal how words have long been used to exclude and divide. What is different now is the speed, scale, and algorithmic spread of online communication.

Bringing these perspectives together shows that criminal law by itself is not enough. Behavioural theories explain how harmful speech spreads in complex ways that punishment cannot easily stop. Regulatory theories show that platforms structure behaviour and must be included in solutions. Normative theories confirm that protection of dignity and equality provides a clear reason for intervention. Historical examples remind us that hate speech adapts to new forms, and the digital environment makes the challenge more urgent. This combined framework supports the case for moving from a narrow criminal law approach to a duty of care model that is systemic, proactive, and consistent with core values of dignity and equality.

Taken together, behavioural, regulatory, and normative theories provide a multi-layered framework for analysing online hate speech. Behavioural theories explain why users engage in harmful expression; regulatory theories reveal how platform design and governance shape these behaviours; and normative theories clarify why such speech undermines dignity, equality, and democratic participation. This integrated approach underpins the thesis's central argument: that criminalisation alone is inadequate, and that a statutory duty of care, grounded in behavioural insight, responsive regulation, and substantive equality, offers a more effective and principled response for New Zealand. Unlike existing analyses that examine psychological, regulatory, or normative theories in isolation, this thesis integrates the three into a single evaluative framework. This framework offers a distinctive contribution by providing a

structured lens for assessing how online hate speech arises, how it is shaped by platform dynamics, and why it warrants principled regulatory intervention in New Zealand. Chapter 3 applies this integrated framework to examine how online hate speech and harmful content manifests in practice and how platform design amplifies harm.

Chapter 3: Social Media, Algorithmic Risks, and the Regulation of Hate Speech and Harmful Content

3.1 Introduction

This chapter examines the digital environment in which online hate speech develops and spreads. It focuses on the role of algorithms and platform design in shaping how users interact, communicate, and encounter harmful content. Chapter 2 outlined the theoretical foundations for this analysis, showing that while freedom of expression is central to democratic life, it is not absolute and must be balanced against dignity, equality, and inclusion.[173] Building on that framework, this chapter explores the technological conditions that influence speech online in how design choices and algorithmic systems can amplify hostility, create echo chambers, and make it difficult to distinguish between free expression and harm. It might be said that before any legal or regulatory response can be effective, it is necessary to understand the technological dynamics that make online hate speech so pervasive.

Social media platforms have changed how fast and how far harmful content can spread. Recommendation systems, instant sharing tools, and easy access for users mean that hate speech can reach many people very quickly. These same systems also enable the circulation of

[173] Refer to Chapter 2, sections 2.3-2.4 (behavioural, normative, and legal theories of free speech and hate speech).

misinformation and disinformation. False or misleading content often interacts with hate speech, reinforcing stereotypes or deepening social divisions. In this sense, hate speech and mis/disinformation are part of the same ecosystem of online harm, even if their legal definitions differ. At the same time, the online environment makes regulation more difficult. Public and private spaces often overlap, content moves easily across borders, and laws from one country are hard to enforce online.[174]

The discussion begins with an analysis of how major platforms such as Meta, Twitter/X, and TikTok influence the spread of hate speech. Their design choices and content rules can reduce harm but also amplify it. The chapter then considers the technical and policy challenges of detecting harmful content, followed by problems of algorithmic suppression and discrimination. Throughout the Chapter, it recognises that misinformation and hate speech are closely connected in practice, as both rely on algorithmic systems that reward attention rather than accuracy and civility. Together, these sections will give a picture of the real challenges in applying the theoretical principles from Chapter 2.[175] They highlight the tension between protecting freedom of expression and keeping online spaces safe and inclusive. The next section turns to algorithmic design and its role in increasing the visibility and influence of harmful speech.

3.2 Social Media and the Amplification of Hate Speech

Social media platforms are now central to digital communication, political discourse, and public debate. However, their engagement-driven design has also facilitated the spread of

[174] Refer to Chapter 2, section 2.5 (regulatory theory: Lessig, Murray, Reidenberg).
[175] Refer to Chapter 2, sections 2.3-2.4 (normative principles of dignity, equality, and proportionality).

harmful content including hate speech, misinformation, and extremist rhetoric. Initially created to connect people, these platforms now operate on engagement-driven and data-monetisation models that amplify polarising content and fuel ongoing debates over platform liability and regulatory intervention.[176] The following examples are used illustratively to demonstrate how platform amplification operates in practice. The analytical claims advanced here is grounded in existing scholarly literature on online disinhibition, platform governance, and digital harm.

Behavioural theories help explain why these dynamics are so powerful. The Online Disinhibition Effect shows how digital environments lower social restraints, making users more willing to post hostile or aggressive content.[177] When such content is promoted by algorithms that reward outrage or sensationalism, it reaches wider audiences and becomes normalised in online culture. Similarly, Social Learning Theory highlights how repeated exposure to harmful material can encourage imitation.[178] When users see hate speech gaining likes, shares, or approval, they may replicate or reinforce the behaviour, especially within tightly connected online communities.

Major platforms illustrate this risks. Meta's recommendation systems have been criticised for driving polarising content higher in user feeds. On X/Twitter, the retweet and trending features accelerate the spread of controversial posts, often without context. TikTok's "For You" page, powered by opaque algorithmic curation, has also been linked to the rapid promotion of

[176] Gary Drenik "Unveiling "X": The Implications Of Twitter's Bold Rebranding Move" Forbes (8 September 2023) <https://www.forbes.com/sites/garydrenik/2023/09/08/unveiling-x-the-implications-of-twitters-bold-rebranding-move/?sh=455aa9d72ff2>.
[177] Refer to Chapter 2, section 2.2 (behavioural theory: Online Disinhibition Effect).
[178] Refer to Chapter 2, section 2.2 (behavioural theory: Social Learning Theory).

discriminatory material.[179] In each case, amplification is not neutral but shaped by the architecture of the platform. The same amplification systems that reward outrage and emotional engagement also provide fertile ground for mis/disinformation. False narratives spread through identical algorithmic channels, often entangled with hate speech or extremist content. This overlap complicates both detection and regulation, as measures aimed at curbing one form of harm can inadvertently affect the other.

The effect is that harmful speech is not only more visible but also more influential. By rewarding content that provokes strong emotional reactions, platforms can magnify the social harm identified in Chapter 2: damage to dignity, erosion of equality, and the exclusion of targeted groups from public discourse.[180]

Criticism of major platforms shows how amplification connects with governance choices. Facebook and Instagram, for example, have faced scrutiny for community guidelines that allow harmful content to be algorithmically promoted rather than reduced. Twitter/X shifted its policies sharply after Elon Musk's acquisition, moving toward what has been called "free speech absolutism." This sparked debate about whether platforms should prioritise maximum expression or accept responsibility for the harms that follow. TikTok's design, based on highly personalised algorithmic feeds, creates further challenges because extremist or misleading material can be widely circulated within short periods of time. These examples highlight how amplification is never neutral but shaped by corporate policies and design. While their

[179] Karen Ho "The Facebook whistleblower says its algorithms are dangerous. Here's why." MIT Technology Review (5 October 2021).
< https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/>
[180] Refer to Chapter 2, section 2.3 (normative theories of dignity and equality).

community guidelines (examined further in Chapter 5) shape what content is permitted, here the focus is on how algorithmic design intensifies harmful speech.

Public controversies show how these dynamics play out in practice. Israel Folau's social media posts on religious discrimination, and Donald Trump's use of Twitter to spread xenophobic messages, both demonstrate how individual users can harness platform features to broadcast harmful ideas at scale.[181] In Aotearoa New Zealand, the political advocacy group Hobson's Pledge provides a striking case. [182] The group campaigns against what it describes as "race-based policies," but its messaging often targets Māori initiatives such as co-governance, Te Reo Māori promotion, and constitutional recognition of Te Tiriti o Waitangi. Through coordinated social media activity, including paid advertising, Hobson's Pledge spreads a narrative that frames Māori as beneficiaries of unfair privilege. [183] While this type of content may not always meet the legal threshold for hate speech, it sits within the same communication space. It tends to normalise hostility toward protected groups and shows how online systems allow borderline speech to circulate widely and shape public attitudes.

This case illustrates the limits of New Zealand's legal framework. Under section 14 of the New Zealand Bill of Rights Act 1990, such speech is protected as political expression.[184] Yet, as Waldron argues, the harm of hate speech is not limited to incitement to violence. It also erodes dignity, weakens the assurance of equal status, and narrows participation in public life.[185] The

[181] Andrew Tobin and Jon Erbacher "The Israel Folau saga: A simple case of failure to comply with an employment contract or is there more?" (20 May 2019) < https://www.hopgoodganim.com.au/news-insights/the-israel-folau-saga-a-simple-case-of-failure-to-comply-with-an-employment-contract-or-is-there-more/>.
[182] Barendze, above n 165, at 1-3 and 61-65.
[183] Barendze, above n 165, at 67.
[184] New Zealand Bill of Rights Act 1990, s 14.
[185] Waldron, The Harm in Hate Speech, discussed in Chapter 2, section 2.3.

example of Hobson's Pledge shows how online communication can fall short of the legal definition of hate speech but still carry similar social effects.

These examples underline the central regulatory challenge: how should platforms balance the protection of free expression with the need to prevent social harm? This guiding question frames the case studies that follow, beginning with Meta (Facebook and Instagram), Twitter/X, and TikTok.

3.2.1 Meta (Facebook and Instagram)

Meta's dominance in global digital communication makes its approach to hate speech one of the most consequential experiments in private regulation. Yet its policies are shaped less by human rights principles than by commercial incentives, leaving dignity and equality precariously protected. With more than 2.5 billion users, Meta's policies and technical design influence how people communicate and take part in democracy.[186] Its targeted advertising business model creates incentives for algorithms to amplify sensational or polarising content, often including hate speech, which makes platform governance a central concern for addressing online harm.[187]

The Terms of Service give Meta broad discretion to remove content that creates "regulatory or legal risk".[188] However, the company also allows political hate speech to remain online under

[186] Christopher McFadden "A Brief History of Facebook, Its Major Milestones" (2020) <https://interestingengineering.com/history-of-facebook>.
[187] Katherine Herrick "Breaking Things: Origins and Consequences of Racialized Hate Speech on Facebook" (International Studies Honors Projects, Macalester College, 2022) at 39.
[188] Facebook "Terms of Service Section 3.2" (2025) <https://www.facebook.com/terms.php>.

a "newsworthiness" or "public interest" exception".[189] This discretionary power lacks transparency and consistency. From Mill's liberal perspective, this discretion corrodes the conditions for rational debate, since what is removed or retained is decided not by reasoned contestation but by non-transparent corporate authority. Waldron would diagnose an even deeper harm: when hate speech is left online, the dignity of vulnerable groups is systematically undermined, normalising exclusion from public life.[190]

Litigation illustrates how external pressure can shape platform practices. In *Fraley v. Facebook*[191] the settlement required Meta to modify its advertising policies, showing that legal constraints can compel change. However, there has been no equivalent judicial mechanism for hate speech, leaving platform discretion largely unchecked. For jurisdictions such as New Zealand, where statutory duties are absent, this reliance on voluntary initiatives such as the Christchurch Call is fragile or even illusory. In this context, protections cannot truly be said to exist at all, beyond voluntary promises and soft-law commitments. Breyer's proportionality approach helps explain this weakness: without binding duties, voluntary commitments fail to secure balanced protection between free expression and the prevention of harm.[192]

The leaked Facebook Papers highlighted these problems. Internal reports admitted that detection systems for hate speech were limited outside English-speaking countries, with severe consequences in Myanmar and India, where Facebook was linked to communal violence.[193] These failures demonstrate the risks of over-reliance on self-regulation in multilingual and

[189] Al Jazeera, "Critics say Twitter treats hate speech as being 'public interest'" (16 October 2019) Al Jazeera < https://www.aljazeera.com/ajimpact/critics-twitter-treats-hate-speech-public-interest-191016225423140.html >.
[190] Elizabeth Dubois and Anna Reepschlager "How harassment and hate speech policies have changed over time: Comparing Facebook, Twitter and Reddit (2005-2020)" (2023) 16 Policy Internet 523 at 530.
[191] *Fraley v. Facebook, Inc.* 830 F. Supp. 2d 785 (N.D. Cal. 2011).
[192] Evelyn Douek "Governing Online Speech: From "Posts-As-Trumps" To Proportionality And Probability" (2021) 121 Columbia Law Review 759 at 767; see also Chapter 2, sections 2.3-2.4.
[193] Evelyn Douek "Governing Online Speech: From "Posts-As-Trumps" To Proportionality And Probability" (2021) 121 Columbia Law Review 759 at 767.

multicultural contexts. In Aotearoa New Zealand, Māori, Pacific, and migrant communities remain vulnerable to digital exclusion and racialised abuse. This reality underscores Meiklejohn's civic republican principle that equal participation in democratic discourse cannot be left to the goodwill of platforms. If Meta cannot secure equality of participation, then regulation is not optional but essential.[194] More than this, the New Zealand experience shows that protecting speech requires a proactive state role: to translate abstract commitments to dignity and equality into enforceable safeguards against the harms produced by global platform design.

3.2.2 Twitter/X and the Free Speech Debate

What does "free speech absolutism" mean in practice, and what happens when a platform of Twitter's scale adopts it as policy? Twitter, now rebranded as X following Elon Musk's 2022 acquisition, has undergone significant changes in content moderation policies.[195] Musk has publicly declared his vision of Twitter as a "free speech absolutist", invoking a libertarian reading of Mill that equates free speech with minimal interference.[196] Yet Mill's defence of liberty presupposed conditions for rational debate, not a marketplace skewed by algorithmic amplification.[197]

Under Musk's leadership, the company has reduced enforcement capacity, dissolved trust and safety teams, reinstated banned accounts, including those previously suspended for hate speech and incitement, and removed longstanding moderation policies related to COVID-19

[194] Refer to Chapter 2, section 2.5.
[195] Drenik, above n 176.
[196] Drenik, above n 176.
[197] Refer to Chapter 2, section 2.5 (Mill).

misinformation and targeted abuse.[198] Musk also abolished the Civic Integrity Policy, which previously prohibited misleading claims intended to suppress participation in civic processes, such as elections.[199]

These changes reflect a shift from "content governance" to "content permissiveness", in which platform architecture no longer mitigates harm but facilitates it.[200] Notably, X has defunded key safety tools, reduced transparency reporting, and reversed bans on controversial figures such as Andrew Tate, Kanye West, and Donald Trump; all of whom had been previously deplatformed for hate speech or inciting harm.[201] Musk has also publicly criticised and, at times, suspended the accounts of journalists and researchers who monitor hate speech trends.

Waldron's perspective highlights why this shift is harmful: permissiveness corrodes the dignity and equal status of minorities, creating what he calls a "hostile environment" for public participation.[202] Meiklejohn would add that democratic discourse cannot flourish if intimidation and abuse silence certain voices, meaning Musk's "absolutism" paradoxically undermines the conditions for genuine free speech.[203]

Before Musk's acquisition, Twitter had taken a more active, if imperfect, approach to content moderation. It used contextual labelling and algorithmic demotion to flag harmful tweets, particularly around misinformation, electoral disinformation, and hate speech. It also

[198] Farhan Asif Chowdhury, Dheeman Saha, Md Rashidul Hasan, Koustuv Saha , Abdullah Mueen "Examining Factors Associated with Twitter Account Suspension Following the 2020 U.S. Presidential Election." (2021) Arxiv <https://doi.org/10.48550/arXiv.2101.09575>
[199] Adrian Kopps Two years after the takeover: Four key policy changes of X under Musk" (28 October 2024) The Humboldt Institute for Internet and Society <https://www.hiig.de/en/policy-changes-of-x-under-musk/>
[200] Dan Valeriu Voineo, "Taking Over Twitter - Balancing Free Speech And Content Moderation" (2022) 8 Annals of the University of Craiova for Journalism, Communication and Management 139 at 140.
[201] Kopps, above n 199.
[202] Refer to Chapter 2, section 2.5 (Waldron).
[203] Refer to Chapter 2, section 2.5 (Meiklejohn).

developed initiatives like "Birdwatch," a crowdsourced fact-checking system.[204] However, the platform's architecture and policy framework have now dramatically changed. [205]

X's trajectory under Musk illustrates the dangers of deregulated or purely profit-driven content moderation. The removal of protective mechanisms has coincided with a documented rise in hate speech, antisemitic content, and transphobic slurs on the platform. As Waldron argues, such environments corrode dignity and equal status, narrowing participation in public life and excluding vulnerable communities from democratic discourse.[206]

Even before Musk's ownership, however, Twitter's enforcement was inconsistent. The case of Donald Trump's suspension in 2021 highlights the platform's discretionary power. Despite years of xenophobic and inflammatory rhetoric, Trump was only banned after the January 6th U.S. Capitol riots. From Mill's perspective, this form of selective intervention undermines rational debate because decisions are shaped by political optics rather than principled, consistent application of standards.[207]

From a legal perspective, the question remains whether platforms like Twitter/X should be considered publishers (and therefore liable for hosted content) or neutral intermediaries under Safe Harbour provisions. In New Zealand, the Harmful Digital Communications Act 2015 (HDCA) does not impose a duty of care on platforms, meaning harmful content can persist unless reported by users. This gap can be explained through Breyer's proportionality analysis:

[204] David La Barbera, Eddy Maddalena, Michael Soprano, Kevin Roitero, Gianluca Demartini, Davide Ceolin, Damiano Spina and Stefano Mizzaro "Crowdsourced Fact-checking: Does It Actually Work?" (2024) 61 Information Processing & Management < https://doi.org/10.1016/j.ipm.2024.103792>
[205] Kopps, above n 199.
[206] Refer to Chapter 2, section 2.5 (Waldron).
[207] Refer to Chapter 2, section 2.5 (Mill).

without binding duties, the balance tilts heavily toward platform autonomy, with insufficient safeguards against the harms of online speech.[208]

By contrast, the European Union's Digital Services Act (DSA) and the United Kingdom's Online Safety Act impose proactive obligations on platforms to reduce illegal and harmful content. These frameworks reflect Meiklejohn's principle that equal participation in democratic discourse requires structural protections against environments that silence vulnerable voices. [209] Compared with New Zealand's reliance on voluntary compliance, these statutory models better secure both free expression and the conditions of equality that make democratic participation meaningful. In this sense, Musk's "absolutism" offers an answer to the question posed at the start: free speech without safeguards is not freedom at all, but a distortion that corrodes dignity and democracy alike.

3.2.3 TikTok: Algorithmic Amplification and Content Governance

What happens to free speech and harm when algorithms, not human choice, determine what billions of people see? TikTok's rise to global prominence has introduced new regulatory concerns, particularly regarding algorithmic amplification of harmful or misleading content. Unlike Facebook or Twitter, whose platforms primarily deliver content through networks of social connections (such as "friends" or "followers"), TikTok relies almost entirely on a content-based recommendation engine.[210] This means that rather than seeing posts from people a user has chosen to follow, TikTok's "For You" feed is populated by videos selected through machine learning models based on watch time, engagement rates, video content, and user

[208] Refer to Chapter 2, section 2.5 (Breyer).
[209] Refer to Chapter 2, section 2.5 (Meiklejohn).
[210] Mansoor Iqbal "TikTok Revenue and Usage Statistics" (18 April 2024) <https://www.businessofapps.com/data/tik-tok-statistics/>.

behaviour patterns. This architecture enables virality without accountability: a user can encounter sensationalist, harmful, or hateful content from complete strangers within seconds of opening the app.

This difference in content delivery heightens regulatory challenges, as content can go viral without the network-based checks present on other platforms. This depersonalised feed structure has been linked to the rapid spread of hate speech, disinformation, and harmful trends, particularly among young users, raising questions about how existing regulatory models, which often rely on user reporting or network moderation, can respond effectively.

In 2020, India banned TikTok under section 69A of its Information Technology Act 2000, citing concerns over national security and the platform's failure to adequately moderate harmful content.[211] Similarly, the United States, debates over banning or restricting TikTok have centred on data privacy and national security, particularly due to TikTok's ownership by the Chinese company ByteDance and the potential for state-directed data access.[212] While these concerns primarily relate to data sovereignty, they also expose broader legal challenges posed by foreign-owned, algorithmically driven platforms like questions around enforcement jurisdiction, the extraterritorial reach of domestic laws, and the ability of regulators to ensure compliance from platforms with opaque decision-making systems.

[211] WARC, "TikTok India forecasts 'at least' 50% user growth in 2020" (2020) <https://www.warc.com/newsandopinion/news/tiktok-india-forecasts-at-least-50-user-growth-in-2020/en-gb/43327> and Deepa Christopher "India - Chinese Apps banned as border tensions rise" (8 July 2020) Linklaters <https://www.linklaters.com/en/insights/blogs/digilinks/2020/july/india---chinese-apps-banned-as-border-tensions-rise>

[212] Courtnet Lawton "TikTok ban in the United Sates: A necessary precaution or a misstep" (29 Jan 2025) < https://policyreview.info/articles/news/tiktok-ban-united-states-necessary-precaution-or-misstep/1822> and PBS News "Supreme Court upholds TikTok ban if not sold by Chinese, Trump has promised a solution" (17 January 2025) PBS News < https://www.pbs.org/newshour/politics/supreme-court-upholds-tiktok-ban-if-not-sold-by-chinese-trump-has-promised-a-solution>

The consequences are evident. Harmful content, from racist memes to disinformation campaigns, can achieve mass reach before user reporting or traditional moderation mechanisms can respond. This raises acute problems for frameworks like New Zealand's Harmful Digital Communications Act 2015, which is reactive and user-driven, not systemic. As Suzor and Gorwa argue, such "black box" systems conceal how content is promoted and make it impossible for regulators, or users, to interrogate why harmful material thrives.[213]

The risks are amplified by TikTok's global ownership and governance, resulting in the aforementioned ban by India and the U.S. debates questioning data sovereignty and Chinese state access.[214] These geopolitical moves highlight how algorithmic governance collides with questions of jurisdiction, enforcement, and the reach of domestic law.

From a theoretical perspective, Mill would view TikTok's algorithm as distorting the conditions of rational debate: instead of a marketplace of ideas, the feed is a marketplace of attention, skewed toward outrage. Waldron's dignity-based account shows how amplification of hate speech corrodes equal status and creates hostile environments for minorities.[215] Breyer's proportionality analysis explains why regulatory intervention is needed: without transparency duties, the balance tilts toward platform autonomy at the expense of public

[213] Nicolas Suzor, "Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms" (2018) 4(3) *Social Media + Society* 1 at 5 and Robert Gorwa, "The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content" (2019) 8(2) Internet Policy Review 1 at 10

[214] WARC, "TikTok India forecasts 'at least' 50% user growth in 2020" (2020) https://www.warc.com/newsandopinion/news/tiktok-india-forecasts-at-least-50-user-growth-in-2020/en-gb/43327 and Deepa Christopher "India - Chinese Apps banned as border tensions rise" (8 July 2020) Linklaters <https://www.linklaters.com/en/insights/blogs/digilinks/2020/july/india---chinese-apps-banned-as-border-tensions-rise>.

[215] Refer to Chapter 2, section 2.4 (Waldron).

safety.[216] And Meiklejohn reminds us that democratic discourse depends on structural protections against environments that silence certain voices.[217]

In this sense, the EU's Digital Services Act is notable because it regulates not just content but also the infrastructure of dissemination. Its provisions on algorithmic transparency, systemic risk assessments, and external audits for "very large online platforms" like TikTok represent an attempt to align platform governance with democratic values. By contrast, New Zealand's reliance on voluntary compliance leaves systemic risks untouched.

TikTok therefore exemplifies why regulation must move beyond individual takedowns to systemic accountability. The question is no longer only what content circulates online, but *how* and by *whose* design, it is amplified. These amplification dynamics underline why detection systems have become central to governance debates; if platforms cannot prevent harmful content from going viral, the next question is whether technology can reliably identify and filter it. Taken together, these platform features and incentives create a single, system-level effect: amplification of group-based hostility and suppression of vulnerable voices. This cumulative impact, rather than any one platform's policy, is what motivates the duty-of-care model developed in Chapter 6.

3.3 Detecting Hate Speech: Technical Limits and Normative Stakes

[216] Refer to Chapter 2, section 2.5 (Breyer).
[217] Refer to Chapter 2, section 2.3 (Meiklejohn).

Social media networks operate through algorithmic systems that structure how speech circulates.[218] Understanding these systems is essential for regulating hate speech online. Artificial intelligence (AI) provides the technological foundation for most automated content-moderation tools used to detect online hate speech. It may be described in general terms as computer systems that perform tasks requiring human-like reasoning or perception. AI is often grouped into three broad types:

"…

a) Strong: this level of AI entails the capability for independent thought and reasoning;
b) Weak: this level of AI is designed to simulate human thinking but lacks the capacity for human-like reasoning; and
c) Narrow: at this level, AI programs are tailored to perform specific and well-defined tasks."[219]

In practice, only narrow AI is relevant to hate-speech detection. Detection systems rely mainly on two subfields of AI - Natural Language Processing (NLP) and Machine Learning (ML) - to identify and classify harmful content. Modern models such as Bidirectional Encoder Representations from Transformers (BERT), are trained on large, labelled datasets to recognise context, sarcasm, and evolving coded language, improving on earlier keyword-based filters. Yet as Mill would caution, crude keywords blocking risks supressing legitimate debate whereas, under-detection leaves minorities vulnerable, as Waldron emphasises.

Accuracy depends heavily on the quality and diversity of training data, inclusion of multiple languages, and the ability to adapt to changing online speech patterns. Without human

[218] Pablo Nicolás Terevinto and others "A Framework for OSN Performance Evaluation Studies" in Tansel Özyer and Reda Alhajj (eds) *Machine Learning Techniques for Online Social Networks* (Cham, Springer International Publishing, 2018) at 47.
[219] Michael Legg and Felicity Bell *Artificial Intelligence and The Legal Profession: A Primer* (University of New South Wales, Sydney, 2017) at 2.

oversight, detection systems risk becoming automated arbiters of speech, legitimising errors that silence some voices while leaving others unprotected.

Natural Language Processing (hereinafter "NLP") enables computers to interpret human language and is already used in translation tools, grammar checkers and voice assistants.[220] These familiar examples show both the potential and the limits of applying language models to the contested domain of hate speech regulation.[221]

Machine Learning (hereinafter "ML") builds on human-defined data inputs and rules, allowing systems to learn patterns and make predictions.[222] It includes supervised, unsupervised, semi-supervised, and reinforcement learning.[223] These techniques enable large-scale content classification but still struggle with contextual or culturally specific expressions: a serious issue in multilingual settings such as Aotearoa New Zealand.

A subset of ML known as Deep Learning (DL) uses artificial neural networks inspired by the human brain to process data through multiple layers. DL has improved accuracy in detecting complex linguistic patterns but remains limited by the quality of its training datasets.[224] Responding to the rise of online hate, researchers such as Mozafari et al. developed transfer-learning approaches using pre-trained language models like BERT.[225] This method allows

[220] Dr. Michael J. Garbade "A Simple Introduction to Natural Language Processing" (2018) Becoming Human <https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32>.
[221] Garbade, above n 220.
[222] Brandon Reagen and others "Deep Learning for Computer Architects" (Springer International Publishing AG, Switzerland, 2022) at 17.
[223] Ed Burns "Definition: Machine Learning" (September 2023) Search Enterprise AI <https://searchenterpriseai.techtarget.com/definition/machine-learning-ML>.
[224] Jason Brownlee "What is Deep Learning?" (2020) <https://machinelearningmastery.com/what-is-deep-learning/>.
[225] Reza Farahbakhsh, Marzieh Mozafari, and Noël Crespi "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media" (paper presented to the Eighth International Conference on Complex Networks and Their Applications, France , 2019).

detection systems to learn from vast existing text corpora, improving accuracy even when labelled hate-speech data is scarce.

Technical systems for detecting hate speech also raise normative questions. From Mill's view, over-blocking threatens the "marketplace of ideas," while Meiklejohn would argue that meaningful discourse may require limits on corrosive abuse.[226] Waldron reminds us that misclassifying targeted hate as mere offence denies minorities equal respect, and Breyer's proportionality framework calls for balanced regulation that protects both speech and dignity.[227] In this sense, the success of detection systems should be judged not only by their technical accuracy but also by whether they promote democratic legitimacy and civic equality online. Detection, however, represents only one side of algorithmic governance. Platforms also shape speech through suppression and curation, sometimes reinforcing offline inequalities. The next section turns to this darker dimension of algorithmic control.

3.4 Platform Discrimination and Algorithmic Suppression

This discussion provides contextual background illustrating how platform dynamics shape the conditions under which hate speech emerges, supporting the later regulatory analysis. In 2020, internal documents revealed that TikTok instructed its moderators to suppress content from users considered unattractive or poor to maintain an "aspirational" image for the platform.[228] This practice illustrates how algorithmic moderation can reflect social hierarchies, even when the stated aim is to enhance user experience. Moderators were told to limit the visibility of people described as having an "abnormal body shape," being "chubby," "too thin," or showing

[226] Refer to Chapter 2, section 2.2.
[227] Refer to Chapter 2, section 2.5.
[228] Sam Biddle, Paulo Victor Ribeiro and Tatiana Dias "Invisible Censorship - TikTok Told Moderators to Suppress Posts by "Ugly" People and the Poor to Attract New Users" (2020) The Intercept <https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/>

“ugly facial looks or deformities.” Similar treatment applied to users posting from “rural” or “slum” environments with “cracked walls” or “disreputable decorations."[229]

The same leaked guidelines also instructed staff to restrict certain political or identity-based content. Moderators were told to censor livestreams that criticised governments or were said to harm “national honour (sic),” and to flag posts by LGBTQ+ users or those considered overweight as “special” categories requiring additional control.[230] These rules demonstrate that suppression is not only about hate speech or harm prevention, but also about shaping what kinds of identities and viewpoints remain visible online.

What began as a claim of “safety” effectively turned into a form of hidden suppression that silenced marginalised voices. Research by Ungless, Markl, and Ross supports this conclusion.[231] Their study of over 600 UK-based TikTok users found that LGBTQ+ and other minority creators frequently experienced perceived censorship even when following community guidelines. This evidence highlights a key tension: platforms justify suppression in the name of protection, yet these same practices chill expression and undermine trust. These findings reveal the deeper regulatory problem at the centre of this chapter: without transparency and accountability, platform moderation can reproduce existing inequalities while claiming to prevent harm.

[229] The Guardian “TikTok 'tried to filter out videos from ugly, poor or disabled users'” (17 March 2020) <https://www.theguardian.com/technology/2020/mar/17/tiktok-tried-to-filter-out-videos-from-ugly-poor-or-disabled-users>
[230] Biddle and others, above n 228 and Chris Köver and Markus Reuter “TikTok curbed reach for people with disabilities” (2019) <https://netzpolitik.org/2019/discrimination-tiktok-curbed-reach-for-people-with-disabilities/>
[231] Eddie L. Ungless, Nina Markl, and Björn Ross, “Experiences of Censorship on TikTok Across Marginalised Identities” (2024) arXiv:2407.14164 <https://arxiv.org/abs/2407.14164>

Waldron's idea of dignity helps explain why algorithmic suppression matters, but as Evelyn Douek observes, the greater challenge lies in holding platforms accountable for how moderation decisions are made. Together, these perspectives show that online visibility is not only about individual expression but also about the systems that determine whose voices are heard in digital spaces.[232]

This moderation strategy raises significant free speech concerns because it reduces the visibility of certain groups based on appearance, identity or socioeconomic background. Brown and Sinclair argue that such practices constitute hybrid-attacks, which are discriminatory acts that combine aesthetic or socio-spatial judgments with speech or design decisions to undermine the dignity of particular groups.[233]

The intersection of aesthetic and socioeconomic discrimination in content moderation underscores the need for transparent and equitable platform policies. Without accountability, such practices can reinforce societal inequalities and hinder the inclusive nature of digital spaces. Here, the limits of traditional hate speech law become clear. By focusing on overt incitement, legal doctrine often overlooks structural exclusion. This gap is particularly relevant for New Zealand; neither the Human Rights Act 1993 nor the Harmful Digital Communications Act 2015 even recognises algorithmic suppression, leaving structural exclusion legally invisible.[234]

[232] Refer to Chapter 2, section 2.3 and Evelyn Douek "Governing Online Speech: From "Posts-As-Trumps" To Proportionality And Probability" (2021) 121 Columbia Law Review 759 at 767.
[233] Alexander Brown and Adriana Sinclair, Hate Speech Frontiers: Exploring the Limits of the Ordinary and Legal Concepts (Cambridge University Press, 2023) at 89.
[234] Refer to Chapter 4.

TikTok's internal labelling system reportedly classified LGBTQI-related content under the designation “Risk 3.4,” throttling visibility or triggering geo-blocking in some jurisdictions.[235] While justified as compliance with local law, this practice illustrates a tension between global governance and jurisdiction-specific censorship. From a normative perspective, the problem is not only censorship itself but the lack of accountability for how platform design decides whose voices are heard.

This ambivalence is illustrated by the Black Lives Matter (BLM) movement. Originating in the United States as an online response to the killing of George Floyd, BLM rapidly evolved into a global phenomenon, spreading to countries including New Zealand through hashtags, video content, and solidarity campaigns.[236] Here, the same algorithms that silenced some voices enabled others to organise and amplify demands for racial justice.[237]

Henderson notes that while platforms amplify democratic engagement, they also blur the boundaries between hate speech and other forms of harmful online content.[238] Yet, much of the existing scholarship focuses either on the empowering potential of digital platforms or on their role in spreading hate speech. This binary overlooks how the same infrastructure produces both outcomes simultaneously. This duality reinforces the central claim of this thesis: New Zealand’s legal framework, while protective of speech rights, lacks mechanisms to ensure platform accountability or redress systemic harm.

---

[235] Markus Reuter and Chris Köver “Cheerfulness and censorship” (2019) <https://netzpolitik.org/2019/cheerfulness-and-censorship/>

[236] Aleem Maqbool "Black Lives Matter: From social media post to global movement" (10 July 2020) BBC News < https://www.bbc.com/news/world-us-canada-53273381>

[237] Ruben Enikolopov and others "Social Media and Protest Participation: Evidence From Russia" (2020) 88 Econometrica 1479 at 1482.

[238] Jennifer Jacobs Henderson "New Boundaries of Free Speech in Social Media" in Daxton R Stewart (ed) *Social media and the law : A guidebook for communication students and professionals* (2nd ed, Routledge, New York 2017) at 1.

As Lessig's regulatory theory shows, online speech is shaped by law, norms, markets, and code. Meaningful reform requires what Lessig calls "synergic regulation," where legal interventions (such as hate speech law or duty of care obligations), platform design (e.g. friction-inserting tools, demotion of harmful content), public education, and commercial incentives operate together rather than in isolation.[239]

In light of the challenges posed by online hate, it is essential to assess whether the existing legal framework in New Zealand adequately assigns accountability (whether to individual users, platforms, or both). Despite voluntary efforts by social media companies to introduce content moderation systems and reporting tools, these strategies often function as short-term palliatives rather than systemic solutions. As Binny et al. argue, current counter-hate strategies resemble "band-aids on deep cuts": fragmented, inconsistent, and inadequate without stronger institutional frameworks and cross-sector collaboration. [240] These shortcomings highlight the need to move beyond voluntary measures toward more coordinated models of governance. These weaknesses point to a larger regulatory dilemma: how can New Zealand, and the global community more broadly, develop frameworks that move beyond fragmented, voluntary responses? The concluding section turns to this challenge, framing it through the Christchurch Call and the theory of dynamic regulation.

3.5 Mis/disinformation

Mis/disinformation does not always meet the threshold of hate speech, yet the two frequently intersect. Hate speech relies on emotional and group-based division, while mis/disinformation

[239] Lawrence Lessig *Code: And Other Laws of Cyberspace* (Basic Books, New York, 1999) at 164.
[240] Mathew Binny and others "Thou Shalt Not Hate: Countering Online Hate Speech" (2019) 13 Proceedings of the International AAAI Conference on Web and Social Media Full Papers.

provides the factual distortions that make such division appear credible. Together they illustrate how digital platforms amplify harm through both emotional and informational manipulation.

This section situates mis/disinformation within the broader discussion of online harm, focusing on how it interacts with hate speech and the challenges it creates for regulation. Misinformation refers to the unintentional spread of false information, whereas disinformation involves the deliberate creation and distribution of deceptive content. Hate speech, by contrast, targets individuals or groups based on protected characteristics such as race, religion, gender, or ethnicity. Although distinct, misinformation and disinformation fuel the spread of hate speech by distorting public perception, legitimising prejudice, and exacerbating social division.

The link between misinformation and hate speech is particularly evident in moments of political or social crisis. Both misinformation and disinformation can intensify hate speech by reinforcing existing biases and propagating false claims about specific groups.[241] False narratives (including conspiracy theories) have contributed to the radicalisation of individuals and have led to real-world violence, as seen in the Christchurch terrorist attack (2019), where the perpetrator was influenced by online disinformation. During the COVID-19 pandemic, misinformation blaming certain communities for spreading the virus contributed to increased xenophobia and hate speech. [242] These incidents show how false information can create fertile ground for hate-based narratives to grow. Social media platforms, designed around engagement and virality, often accelerate this process by rewarding outrage and emotional reaction rather than accuracy. In this environment, hate speech and misinformation reinforce each other, creating an ecosystem of digital harm that transcends borders.

[241] Sofia Cherici "Mirroring Bias: Online Hate Speech and Polarisation" *Green European Journal* 2021 < https://www.greeneuropeanjournal.eu/mirroring-bias-online-hate-speech-and-polarisation/>.
[242] Peter Billie Larsen and Marjorie Pamintuan "The Human Right to Science: From Fragmentation to Comprehensive Implementation?" (Research Paper, No. 163, South Centre, Geneva 19 August 2022) at 19.

While later chapters will examine regulatory frameworks in detail, it is important to note here that New Zealand's current system remains reactive and focused on individual harm rather than systemic amplification. The following section turns to specific events that reveal how misinformation and hate speech interact in practice, demonstrating the tangible consequences of this convergence.

3.5.1 Consequences and Local Dynamics

The Christchurch terrorist attack (2019) demonstrates how misinformation fuels online hate speech and radicalisation. False narratives about immigration and Islam, widely circulated in far-right online spaces, reinforced the attacker's extremist beliefs. In the aftermath, false claims about the event continued to circulate, amplifying fear and division.[243] This demonstrates how real-time falsehoods can intensify the social harms of hate speech and sustain cycles of hostility.

During the COVID-19 pandemic, the spread of health-related misinformation—sometimes described as an "infodemic", fuelled xenophobia, fear, and hostility toward Asian communities. Declining scientific literacy and limited fact-checking capacity deepened public distrust in science and government institutions, making citizens more vulnerable to conspiracy theories and anti-vaccine rhetoric.[244] False claims such as "COVID-19 is just a flu" or "vaccines contain tracking chips" spread rapidly across social media, eroding trust and

[243] Lea Bader and Jochen Bender "What is "fake news" and "hate speech" and how do they work in practice?" Central and Eastern European EDem and EGov Days 342 (March 2022) 17 at 18.
[244] Larsen and Pamintuan, above n 242, at 8.

legitimising prejudice.[245] These narratives demonstrate how misinformation can serve as a bridge between public confusion and targeted hate.

The 2022 Wellington protest, often referred to as the Convoy 2022 occupation, illustrated how online misinformation about COVID-19 mandates evolved into organised offline mobilisation.[246] Anti-vaccine protests not only opposed public-health measures but also became a breeding ground for conspiracy theories and far-right extremism. False narratives that linked government policies to social control fostered hostility toward officials and citizens, turning online rhetoric into real-world confrontation. [247] As extremist ideas such as white nationalism and Christian fundamentalism circulated through online channels, the protests shifted from peaceful expression to violent confrontation, revealing how misinformation can merge with ideological hate. The state's cautious response underscored the tension between maintaining public order and preserving democratic dissent, but the episode left a lasting imprint on public trust and political discourse.

The visit of British anti-transgender activist Posie Parker in 2023 further illustrated the convergence between misinformation networks and hate speech[248]. Online groups that had initially spread COVID-19 conspiracies repurposed their platforms to amplify Parker's campaign, framing it as a "free-speech" movement while vilifying transgender people.[249] The

[245] Bader and Bender, above n 243, at 21.
[246] Digby Werthmuller "Anti-mandate protesters convoy on both North and South Islands" (7 July 2022) 1News < https://www.1news.co.nz/2022/02/07/anti-mandate-protesters-convoy-on-both-north-and-south-islands/>
[247] Eva Corlett and Tess McClure "New Zealand police clash with anti-vaccine protesters at parliament, over 120 arrested" (10 February 2022) *The Guardian* <https://www.theguardian.com/world/2022/feb/10/new-zealand-police-clash-with-anti-vaccine-protesters-during-eviction-operation>
[248] 1News "What are Posie Parker's views and why are they so controversial?" (24 March 2023) 1News < https://www.1news.co.nz/2023/03/24/what-are-posie-parkers-views-and-why-are-they-so-controversial/>
[249] Shanti Mathias, 'Tracking the surge in online anti-trans hate sparked by Posie Parker's visit', date? (5 May 2023) The Spinoff < https://thespinoff.co.nz/internet/05-05-2023/tracking-the-surge-in-online-anti-trans-hate-sparked-by-posie-parkers-visit>.

so-called “Parker effect” showed how misinformation framed as free-speech advocacy can escalate into real-world hostility and public disorder.[250]

Together, these examples underline the urgent need to strengthen media-literacy skills and promote ethical information practices. Encouraging critical engagement with online content is essential to countering misinformation’s role in amplifying hate speech. More broadly, they reveal how digital falsehoods destabilise trust in institutions and deepen social polarisation, setting the stage for examining how New Zealand’s current frameworks address misinformation-driven harm.

Research by the Disinformation Project further illustrates these dynamics. Since 2020, it has tracked how far-right networks in Aotearoa have used social media to spread conspiracy theories and harassment campaigns, particularly during periods of crisis such as lockdowns.[251] Their findings show that misinformation disproportionately targets marginalised and vulnerable communities, deepening distrust and amplifying hostility.[252] Platforms such as Telegram, which lack clear moderation policies, have become key spaces for this activity. Together, these patterns reveal how misinformation does more than distort facts; it corrodes civic trust and entrenches social divisions. The following chapter considers whether New Zealand’s existing frameworks are equipped to respond to these intertwined harms.

[250] Mathias, above n 249.
[251] Sanjana Hattotuwa and Kayli Taylor Kate Hannah "Working Paper, Mis- and disinformation in Aotearoa New Zealand" (The Disinformation Project, 2021) at 7.
[252] Hattotuwa and Hannah, above n 251, at 1.

3.6 Conclusion

The evidence in this chapter shows that the design and operation of online platforms intensify the reach and impact of hate speech. Algorithmic systems that reward engagement also amplify hostility, misinformation, and bias, creating conditions in which harm spreads faster than traditional legal mechanisms can respond. These design choices interact with social and psychological factors, meaning that the harms of online hate speech are not merely legal but structural.

While legal intervention remains necessary, criminal law alone cannot address the systemic features of the digital environment that sustain hate. Effective regulation must therefore begin with an understanding of how algorithms, platform incentives, and information disorders shape communication itself. As Waldron reminds us, dignity and equal status are the benchmarks of legitimate regulation; without them, formal rights risk becoming symbolic, offering recognition without real protection.

In New Zealand, current responses remain reactive and fragmented, often relying on voluntary initiatives and post-hoc enforcement. The next chapter examines the domestic legal framework in greater depth, assessing how far it protects users from online harm and where significant gaps persist. The question that follows is not whether to regulate, but how to ensure that any intervention respects freedom of expression while safeguarding equality, dignity, and civic participation.

Chapter 4: Regulating Hate Speech in New Zealand

This chapter sets out New Zealand's legal framework within the broader category of law and regulation, examining how statutes and case law respond to hate speech and related harms. It concludes that the current approach remains reactive and fragmented, providing limited protection against online hate speech and harmful content, and highlighting the need for a more proactive, duty-based regulatory model. This chapter investigates hate speech in the unique context of New Zealand, focusing on the complex interplay between the current legal frameworks and the notable absence of a dedicated statute addressing hate speech directly. In an increasingly interconnected digital environment, the proliferation of online hate speech raises urgent legal and social questions. These include how to safeguard freedom of expression while protecting individuals and communities from discrimination, vilification, and incitement to harm.

New Zealand, like many liberal democracies, faces the challenge of balancing civil liberties with the need to prevent speech that undermines social cohesion and human dignity. In the absence of a single legislative instrument targeting hate speech, regulation is instead diffused across several statutes, each with different scopes, thresholds, and enforcement mechanisms.

4.1 Background and legal framework

While the central concern of this thesis remains hate speech and harmful content, many of the relevant statutes in Aotearoa are not designed with hate speech in mind but instead address broader categories of harmful or offensive communication. These include the New Zealand Bill of Rights Act 1990, which affirms the right to freedom of expression; the Human Rights Act 1993, which prohibits incitement to racial disharmony; and the Harmful Digital Communications Act 2015, which targets online communications causing serious emotional

distress. Other statutes, such as the Crimes Act 1961, the Summary Offences Act 1981, the Broadcasting Act 1989, and the Films, Videos, and Publications Classification Act 1993, regulate harmful or offensive material more indirectly. This chapter therefore examines not only laws that explicitly address hate speech but also statutes that regulate harmful content more broadly. This wider frame reflects the thesis scope, recognising that online hate speech often intersects with other forms of digital harm, such as harassment, offensive broadcasting, or objectionable publications.

These laws together create a fragmented framework. Some capture aspects of hate speech (such as section 61 of the Human Rights Act), while others address adjacent forms of harmful content (such as online harassment or objectionable publications). Yet there is no coherent statutory scheme that directly and comprehensively addresses online hate speech in its systemic, group-based form. This fragmentation is a central theme of the chapter and underscores the broader argument of this thesis: New Zealand's existing laws provide partial remedies but fail to regulate hate speech effectively in the digital age.

Each of these statutes addresses different aspects of harmful speech (racial incitement, digital harassment, violent threats, or offensive broadcasting), yet there is no unified legal framework that integrates them. This results in a fragmented regulatory landscape in which hate speech may be defined, prosecuted, or even recognised by courts and enforcement agencies. The description of New Zealand's framework as "fragmented" is used here in a functional rather than doctrinal sense. The relevant statutes are internally coherent within their respective domains; however, they were developed to address distinct categories of harm and therefore do not collectively regulate systemic risks arising from platform-scale amplification. In this sense, fragmentation therefore refers to the absence of an integrated regulatory mechanism

capable of addressing cross-platform digital harms rather than to inconsistency within individual statutes. For instance, the threshold for criminal prosecution under the Human Rights Act is high and rarely used, while the HDCA focuses more on harm to individuals than on group-based hate. The Bill of Rights Act provides a constitutional backdrop, but it does not override specific legislative limitations.

This interaction, or lack thereof, between the various laws creates a legal patchwork. The thesis argues that this fragmented model is ill-equipped to respond to the evolving nature of online hate, particularly on social media platforms where anonymity, virality, and algorithmic amplification compound the risk of harm. Understanding how these statutes interact is thus essential to evaluating whether New Zealand's current legal framework offers adequate protection against hate speech in the digital age. While these statutes appear to offer broad coverage, in practice their protections often exist more on paper than in reality.

Without clear statutory duties or consistent enforcement, the framework remains fragmented and underpowered against systemic online harms. New Zealand is unusual in relying on a patchwork of statutes rather than a dedicated hate speech law, a feature explored further below.

4.2 Absence of a Hate Speech Law

Unlike jurisdictions such as the United Kingdom, Canada, and Australia, which have enacted targeted provisions addressing incitement to hatred or vilification, New Zealand disperses regulation across the Human Rights Act, the Harmful Digital Communications Act, the Summary Offences Act, and related legislation.[253] Each statute captures fragments of harmful

[253] Laura O'Connell Rapira and Leroy Beckett "Our Hate Speech Laws" The People's Report on Online Hate, Harassment, and Abuse (2018) < https://actionstation.org.nz/publications >.

speech, but none provides a coherent framework. The absence of a dedicated statute reflects a mixture of political caution and cultural emphasis on expressive freedom.

Following the Christchurch mosque attacks in 2019, then-Justice Minister Hon Andrew Little proposed legislative reform to expand section 61 of the Human Rights Act to cover religion, sexual orientation, disability, and gender identity.[254] He argued that the law, while symbolically important, was "rarely used and too narrow in scope."[255] This led to the Human Rights (Incitement on Ground of Religious Belief) Amendment Bill, which passed its first reading in December 2022 and was referred to the Justice Select Committee. [256] However, in February 2023, Prime Minister Chris Hipkins announced that the bill would be withdrawn. Instead, the government referred the issue to the Law Commission for a comprehensive review on how New Zealand law addresses hate-motivated offending and speech. As of April 2025, this review is ongoing, and no replacement legislation has yet been introduced. From a theoretical perspective, this political polarisation reflects a clash of speech frameworks. Mill's harm principle, dominant in New Zealand political discourse, defends expression until demonstrable harm to others arises. By contrast, Waldron emphasises that hate speech corrodes dignity and social trust even without immediate physical harm, while Fredman argues that failing to restrict such speech undermines substantive equality by normalising marginalisation. The debate therefore reveals not only political disagreement but also divergent conceptions of what counts as "harm."

[254] RNZ "Little plans fast-track review of hate speech laws" (2019) <https://www.rnz.co.nz/news/national/385955/little-plans-fast-track-review-of-hate-speech-laws>.
[255] David Seymour and Andrew Little "Freedom of speech: Do we need to update our Human Rights Act?" (2019) <https://www.stuff.co.nz/national/politics/opinion/113785976/freedom-of-speech-do-we-need-to-update-our-human-rights-act>.
[256] Human Rights (Prohibition of Discrimination on Grounds of Gender Identity or Expression, and Variations of Sex Characteristics) Amendment Bill, Member's Bill, 275-1.

The withdrawal of the 2022 Bill illustrates a persistent reluctance to define hate speech in statutory form, reflecting fear of political backlash rather than principled restraint. This thesis argues that such caution leaves New Zealand with an incoherent framework that neither protects victims effectively nor provides clear guidance to enforcement agencies. While wholesale criminalisation may not be desirable, a measured statutory reform is necessary; especially one that integrates hate-motivated expression within a broader harm-based model and explicitly recognises the structural and cumulative nature of online hostility. In this sense, the Law Commission's review represents an opportunity to move beyond reactive patchwork responses toward a coherent, rights-consistent regime that balances expressive freedom with protection from dignitary and equality-based harms.

In contrast, ACT Party leader David Seymour consistently argued that section 61 unjustifiably restricts free speech.[257] His 2019 Freedom to Speak Bill sought to repeal provisions against insulting or offensive speech, and under the 2024 National-ACT coalition he secured commitments to strengthen academic freedom and expand speech protections. [258] This reflects a liberty-first position that accepts criminalisation of incitement to violence but rejects broader restrictions on offensive or hateful expression.

This divergence reflects a wider philosophical debate: whether the law should criminalise or prohibit certain types of offensive or harmful speech, or whether social regulation (through counter speech and norms) is sufficient. Spoonley supports expanded hate speech protections, especially for online content, arguing that the current legal thresholds are too high and there is

[257] Cameron Walker "David Seymour is Tilting at Free Speech Windmills" (19 June 2019) The Spinoff < https://thespinoff.co.nz/politics/19-06-2019/david-seymour-is-tilting-at-free-speech-windmills>.
[258] John Gerritsen "Government to change free speech rules for universities" (7 December 2024) <https://www.rnz.co.nz/news/political/537143/government-to-change-free-speech-rules-for-universities> and Penny Simmons and David Seymour "Strengthening Free Speech in Universities" (7 December 2024) Beehive.govt.nz <https://www.beehive.govt.nz/release/strengthening-free-speech-universities.>.

a “profound scarcity of successful racial disharmony claims” before the Human Rights Review Tribunal.[259] In contrast, some critics argue that the law should not attempt to define hate speech too broadly, cautioning that doing so may chill legitimate dissent or minority views.[260]

From the perspective developed in Chapter 2, this thesis supports a limited but principled form of statutory regulation. Mill’s harm principle provides an important starting point but underestimates the cumulative and structural nature of online hate. Meiklejohn’s and Waldron’s accounts of speech as a foundation of democratic legitimacy and dignity suggest that law should intervene when expression undermines equal participation or corrodes the assurance of respect within public discourse. Breyer’s proportionality framework offers a way to reconcile such limits with freedom of expression by requiring that restrictions be narrowly tailored and demonstrably justified. Applying these principles, this thesis argues that New Zealand’s current patchwork fails to meet the harm threshold in practice and leaves systemic hostility largely unchecked. A coherent statutory duty, designed around proportionality and dignity, would better align with a democratic theory of speech that protects both open debate and the conditions of equality that make such debate meaningful.

Currently, the Law Commission is reviewing the adequacy of hate speech laws (Law Commission, Hate Crime Consultation Paper; Violence Information Aotearoa, “Submissions open on bill related to hate crime”), but progress has been slow. The Royal Commission into the Christchurch terrorist attack also urged reform, highlighting the link between hate speech

[259] Laura O’Connell Rapira and Leroy Beckett “Our Hate Speech Laws” The People’s Report on Online Hate, Harassment, and Abuse (2018) < https://actionstation.org.nz/publications >.
[260] Ian Bassett "Is hate speech legislation necessary or desirable?"*DayStar Magazine* (Online ed, Auckland, 2005).

and social cohesion.[261] The failure to enact meaningful reform reflects not only political reluctance but also a deeper normative tension in New Zealand law. Mill's harm principle has been given strong weight in judicial and political discourse[262] reinforcing a presumption in favour of expressive liberty. By contrast, theories that emphasise dignity (Waldron) or substantive equality (Fredman) have had little influence on legislative design.[263] The persistence of gaps despite Christchurch reflects New Zealand's reliance on a Mill-ian harm principle in political debate, which prioritises liberty until clear harm arises. Yet, as Waldron's dignity framework stresses, hate speech corrodes equal standing long before violence. Fredman's focus on substantive equality similarly shows how the absence of protection entrenches disparities for Māori, Pasifika and LGBTQ+ groups. The slow pace of reform reflects a deeper normative tension in New Zealand law. Mill's harm principle, strongly embedded in political discourse, prioritises liberty until demonstrable harm arises. By contrast, Waldron's dignity framework and Fredman's substantive equality approach argue that hate speech corrodes equal standing and entrenches disparities well before violence occurs.

This thesis takes the view that New Zealand's continued reliance on a narrow, Mill-ian conception of harm is no longer adequate in an environment shaped by algorithmic amplification and networked hostility. A rights-consistent approach should recognise that harm is not confined to direct injury but also includes the erosion of dignity, belonging, and equal participation in public life. Drawing on Waldron's and Fredman's frameworks, this thesis argues that statutory reform should explicitly incorporate dignity and substantive equality as

[261] Royal Commission of Inquiry Into The Terrorist Attack On Christchurch Mosques On 15 March 2019 "Part 9 - Social cohesion and embracing diversity" (2021) <https://christchurchattack.royalcommission.nz/the-report/part-9-social-cohesion-and-embracing-diversity/hate-crime-and-hate-speech> at 50.
[262] Andrew Geddis "The State of Freedom of Expression in New Zealand: An Admittedly Eclectic Overview" (2008) 11 Otago L Rev 657 at 660-662 and Stephanie Panzic "The Harmful Digital Communications Act 2015: A Deficiency in Defining Harm and an Unintended Limitation on Freedom of Expression" (2015) 21 Auckland U L Rev 218 at 243.
[263] Refer to Chapter 2, section 2.3

guiding principles for regulating hate-motivated speech. Such a shift would not abandon expressive liberty but would situate it within a relational understanding of rights; one that ensures that freedom of expression does not come at the expense of others' ability to participate safely and equally in democratic discourse.

The absence of a dedicated hate speech law is not accidental but reflects this liberty-first model. Marginalised communities are left with narrow, under-enforced provisions such as section 61, while systemic and online hate falls between the cracks.[264] This institutionalised inadequacy makes New Zealand's reliance on general rights protections even more important. The New Zealand Bill of Rights Act 1990, particularly section 14 on freedom of expression, has therefore become the central framework through which debates about hate speech are filtered. Behavioural insights demonstrate that online hate spreads through disinhibition and social learning, yet without a dedicated statutory framework, these harms fall between the cracks. In effect, New Zealand has institutionalised inadequacy: it acknowledges the risks of hate speech but has declined to legislate comprehensively against them.[265]

The stalled reform efforts following Christchurch therefore demonstrate not only political hesitation but also the consequences of treating expressive liberty as an overriding norm. This impasse underscores the need for a more coherent statutory approach, one that aligns freedom of expression with the protection of dignity and equality rather than positioning them in opposition.

[264] Dylan Asafo "The Western Legal Roots of the Christchurch Mosque Shootings: A Colonial Critique of New Zealand's Legal Framework on Racist Hate Speech" (2021) 12 *UC Irvine L Rev* 101 at 17
[265] Refer to Chapter 2, section 2.2.

A coherent statutory duty, designed around proportionality and dignity, would better align with a democratic theory of speech that protects both open debate and the conditions of equality that make such debate meaningful. This legislative gap places greater interpretive weight on the New Zealand Bill of Rights Act 1990, particularly section 14 on freedom of expression.

4.3 New Zealand Bill of Rights Act 1990

The NZBORA is a cornerstone of New Zealand's constitutional framework, outlining the civil and political rights of individuals under domestic law. While New Zealand does not have a single codified constitution, NZBORA functions alongside other constitutional sources, including Te Tiriti o Waitangi, constitutional conventions, and significant statutes such as the Constitution Act 1986, to shape the exercise and limitation of public power.[266]

NZBORA affirms a comprehensive set of civil and political rights, including the rights to life, liberty, peaceful assembly, and freedom from discrimination.[267] Most relevant to this thesis is section 14, which guarantees the right to freedom of expression, defined as "the freedom to seek, receive, and impart information and opinions of any kind in any form." This provision sits at the heart of democratic participation, public debate, and digital communication in New Zealand. However, NZBORA is not supreme law. It does not override other statutes, even if they conflict with its provisions. This is made clear in section 4, which prohibits courts from striking down legislation merely because it is inconsistent with NZBORA. Instead, courts must apply such legislation, even where inconsistency exists.[268] Nevertheless, section 6 directs that,

[266] The Office of the Governor-General "New Zealand's Constitution" (2020) <https://gg.govt.nz/office-governor-general/roles-and-functions-governor-general/constitutional-role/constitution/constitution>.
[267] New Zealand Bill of Rights Act 1990, Part 2.
[268] New Zealand Bill of Rights Act 1990, s 4.

wherever possible, legislation should be interpreted consistently with the rights and freedoms it affirms.[269]

Section 5 sets out the conditions under which rights may be limited: only where those limits are reasonable, prescribed by law, and demonstrably justified in a free and democratic society. This introduces a proportionality test, which has become central to the courts' approach when balancing rights and state interests.[270] In *Moonen v Film and Literature Board of Review*, the Court of Appeal confirmed that limits on s 14 expression must be "demonstrably justified" under s 5 and required a proportionality analysis, preferring rights-consistent interpretations where reasonably open[271]. Moonen shows how proportionality operates under NZBORA, but it also exposes a limit: the test is highly context sensitive and leaves wide discretion, which makes it an uncertain tool against hate speech. From a theoretical perspective, Moonen reflects a Mill-ian tendency to prioritise liberty until harm is demonstrable, but it does not adequately capture Waldron's point that cumulative exposure to hate speech corrodes equal standing. The Court's emphasis on contextual balancing leaves much to judicial discretion, making it ill-suited for addressing systemic or coded forms of online hate speech.

This approach was further developed in *R v Hansen*, where the Supreme Court considered the interaction between ss 4, 5, and 6, offering a structured method for courts to assess whether limitations are justified.[272] In *Hansen*, the Supreme Court grappled with the interpretive method to apply when an enactment conflicts with a right affirmed in the New Zealand Bill of Rights Act 1990 (NZBORA), particularly the presumption of innocence under section 25(c). The Court ultimately confirmed that section 4 prevents courts from invalidating inconsistent

[269] New Zealand Bill of Rights Act 1990, s 6.
[270] New Zealand Bill of Rights Act 1990, s 5.
[271] *Moonen v Film And Literature Board Of Review* [2000] 2 NZLR 9 (CA) at [17].
[272] *R v Hansen* [2007] NZSC 7, [2007] 3 NZLR 1 at [92]-[98].

legislation, while section 5 allows such inconsistencies only where the limitation is "demonstrably justified in a free and democratic society".[273] Tipping J proposed a structured approach: (1) identify whether a right is infringed; (2) assess whether the limit is justified under section 5; (3) if unjustified, determine whether an alternative, rights-consistent interpretation is possible under section 6; and (4) if not, apply the inconsistent statute under section 4.[274]

The Court reiterated that a limitation must pursue an objective of sufficient importance and must impair the right no more than is reasonably necessary to achieve that aim (echoing the Canadian *Oakes* test, but applying it within New Zealand's statutory context).[275] This case thus remains the leading authority on the interpretive methodology for reconciling NZBORA rights with conflicting statutory provisions.

The Supreme Court in *Hansen* refined the proportionality approach and clarified the sequence of analysis and evidential burdens under s 5.[276] *Hansen* underscores judicial deference to Parliament. Without clear statutory direction, courts are reluctant to expand restrictions on expression in hate speech disputes. The structured methodology in *Hansen* echoes Breyer's "active liberty" ideal of promoting democratic participation through proportionate limits. Yet the Court did not grapple with how hate speech undermines participation by silencing minorities. In practice, this shows how a liberty-first approach leaves little space for theories attentive to dignity or substantive equality.

This analysis clarified that while NZBORA affirms important rights, including freedom of expression under section 14, it does not override later inconsistent statutes unless a rights-

[273] *R v Hansen* [2007] NZSC 7, [2007] 3 NZLR 1 at [170]-[180].
[274] *R v Hansen* [2007] NZSC 7, [2007] 3 NZLR 1 at [192].
[275] *R v Hansen* [2007] NZSC 7, [2007] 3 NZLR 1 at [64] citing *R v Oakes* [1986] 1 SCR 103.
[276] *R v Hansen* [2007] NZSC 7, [2007] 3 NZLR 1 at [64].

consistent interpretation is reasonably available. The proportionality tests in *Moonen* and *Hansen* echo Breyer's "active liberty" focus on democratic participation, but NZ courts rarely engage with speech's exclusionary effects. Mill's liberty-based model dominates, leaving little space for frameworks attentive to equality or dignity.

Legal scholars have expressed mixed views on whether this framework provides sufficient protection for expressive freedom. Rishworth notes that the requirement for justified limitation introduces complexity and "judicial restraint," particularly where Parliament has spoken clearly.[277] Similarly, Geddis argues that in controversial areas like hate speech, the courts often defer to legislative choices, which may result in under-enforcement of rights where limits are politically motivated or overly broad.[278]

This judicial and academic landscape has significant implications for hate speech regulation. While NZBORA protects freedom of expression, that protection is subject to lawful and justified limits. Any new legislation that seeks to regulate online hate speech, including on social media, must be drafted with careful attention to section 5, ensuring that any restriction is clear, narrowly tailored, and proportionate. The courts play a pivotal role in reviewing the compatibility of such laws with NZBORA, which positions the Act not only as a safeguard for rights, but as a constitutional framework for evaluating regulatory reform.

NZBORA provides a critical framework for balancing expression and restriction, but the jurisprudence in *Moonen* and *Hansen* illustrates both the promise and the limits of proportionality. The courts' cautious, liberty-centred approach reflects Mill's harm principle more than Waldron's or Fredman's equality-based frameworks, leaving hate speech regulation

[277] Paul Rishworth "Interpreting and Invalidating Enactments Under a Bill of Rights" in Rick Bigwood (ed) The Statute: Making and Meaning (LexisNexis, Wellington, 2004) 251 at 277.
[278] Geddis, above n 262, at 660-662.

politically, rather than judicially, driven. The limits of NZBORA's proportionality framework become clearer when examined against the statutory provisions that do exist. The Human Rights Act 1993 gives a concrete form to anti-discrimination principles, yet its scope and application show how narrow and contested New Zealand's regulation of hate speech remains.

The Court in *Wall v Fairfax* also referred to the interpretive approach established in *Moonen v Film and Literature Board of Review* to reinforce the need for consistency between hate speech restrictions and the New Zealand Bill of Rights Act 1990 (NZBORA). In *Moonen*, the Court of Appeal established a structured framework for determining whether a limitation on expression is justified under section 5 of the NZBORA.[279] This includes identifying whether a right has been limited, whether the limitation is prescribed by law, and whether the objective is sufficiently important to justify the limitation. Crucially, the court must also consider whether the limitation impairs the right no more than is reasonably necessary; a proportionality analysis akin to the *Oakes* test used in Canadian jurisprudence.[280]

4.4 Human Rights Act 1993

The Human Rights Act 1993 (hereinafter "HRA") is a central part of New Zealand's anti-discrimination framework. It promotes equality of opportunity and prohibits both direct and indirect discrimination across a range of protected grounds, including race, ethnicity, gender, sexual orientation, religious belief, and disability. The Act gives practical effect to New Zealand's international obligations, especially under the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD). It also complements civil rights protections under the New Zealand Bill of Rights Act 1990.

[279] *Moonen v Film And Literature Board Of Review* [2000] 2 NZLR 9 (CA) at [17]-[19].
[280] *R v Hansen* [2007] NZSC 7, [2007] 3 NZLR 1 at [64] citing *R v Oakes* [1986] 1 SCR 103.

The HRA applies to everyone in New Zealand, not only to government bodies. It covers private individuals, organisations, and public authorities in areas such as employment, education, and the provision of goods and services.

Section 61 makes it unlawful to publish or distribute written material, or to use words in a public place or broadcast, that are threatening, abusive, or insulting, and are likely to excite hostility against or bring into contempt a group of people based on colour, race, or ethnic or national origins.[281] The test is objective. A person may breach the law even if they did not intend to cause hostility, so long as the words are likely to have that effect.[282]

Complaints under section 61 go first to the Human Rights Commission. If they cannot be resolved, the case can be taken to the Human Rights Review Tribunal. The Tribunal may issue a declaration or award damages, but it cannot impose criminal penalties. Section 131 creates a related criminal offence of inciting racial disharmony, but prosecutions are rare. The provision remains largely symbolic.

New Zealand courts have applied section 61 conservatively. In the leading case of *Wall v Fairfax New Zealand Ltd*, the High Court considered complaints against a series of cartoons published by The Press.[283] Although the Court found the cartoons objectively offensive and racially insensitive, it held they did not meet the high threshold required under section 61 to amount to incitement of racial hostility. This illustrates the difficulty of proving that speech

[281] Human Rights Act 1993, s 61.
[282] *Wall v Fairfax New Zealand Ltd* [2018] NZHC 104.
[283] *Wall v Fairfax New Zealand Ltd* [2018] NZHC 104.

reaches the legal threshold of "exciting hostility or contempt", and also why some advocates see the law as ineffectual in addressing modern forms of hate, particularly online.

Despite various efforts at reform, New Zealand's legal regime remains narrowly focused, covering only racial and ethnic incitement and excluding categories such as religion, sexual orientation, and gender identity. As Geddis notes, courts tend to defer to Parliament in this area, reflecting a strong reluctance to interfere with expressive freedoms unless clear statutory justification exists.[284] Whether such a cautious approach adequately protects vulnerable communities in the digital age remains an open and pressing question.

The rights most relevant to this thesis are those under Part 1A and Part 2 of the HRA, which prohibit discrimination by public bodies and private individuals respectively, and Part 4, which governs the functions and powers of the Human Rights Commission. These include the right to freedom from racial and ethnic discrimination, and the right to protection from incitement to hostility or contempt on racial or ethnic grounds.

The HRA has its origins in the Race Relations Act 1971, which was repealed in 1993 after a comprehensive law reform process to integrate various strands of anti-discrimination law into a more unified structure. Section 61 of the Human Rights Act 1993 ("Racial disharmony") makes it unlawful for any person to publish, distribute, or publicly use words that are threatening, abusive, or insulting and that are likely to excite hostility against or bring into contempt a group of people on the grounds of colour, race, or ethnic or national origins. The provision covers both written and spoken communication, including material published online,

[284] Geddis, above n 262, at 660-662.

and applies regardless of whether the speaker intended to cause hostility.[285] In practice, this means it is illegal to disseminate or broadcast material that fosters racial animosity or prejudice through print, digital media, or public speech, including at meetings or gatherings accessible to the public.[286] The law also applies where a person knew, or ought reasonably to have known, that such material would be made public through newspapers, radio, or television.[287]

The main purpose of this Act is to mitigate and resolve any conduct that may provoke animosity or engender disdain against any collective of individuals on the basis of their colour, race, or ethnic and national backgrounds. To enhance comprehension, the section presents precise definitions. The term "newspaper" refers to a publication that is issued at regular intervals, with a maximum frequency of three months, and encompasses public news, commentary on public news, or primarily consisting of advertisements.[288] The term "publishes or distributes" pertains to the action of disseminating information to either the entire public or specific persons. Finally, the term "written matter" embraces a diverse range of communication modalities, including written text, signs, visible representations, and audio recordings.

Expanding upon the implications of Section 61, this provision essentially prohibits publishing or disseminating content (including online posts) that is threatening, abusive, or insulting and that is likely to encourage hostility or result in contempt toward a group based on race, ethnicity or national origin. The section covers various forms of expression, including “any writing, signs, visible representations, or sound recording.” However, while the statutory language is broad, courts have interpreted these terms narrowly, applying a high threshold for liability.

[285] Human Rights Act 1993, s 61.
[286] Human Rights Act 1993, s 61(1).
[287] Human Rights Act 1993, s 61(2).
[288] Human Rights Act 1993, s 61(2).

This raises particular challenges in the digital age, where offensive content may be widely shared but not necessarily meet the legal standard for incitement. The interpretation of terms such as "visible representation" and "likely to excite hostility" has been considered in *Wall v Fairfax New Zealand Ltd* [2018] NZHC 104[289], which still remains the leading case on how section 61 operates in relation to online and visual content.

In brief, this case had its origins in the Human Rights Review Tribunal and revolved around a legal dispute involving several parties. Louisa Hareruia Wall acted as the Plaintiff, while the Defendants Fairfax New Zealand Ltd (First Defendant), The Marlborough Express (Second Defendant), The Christchurch Press (Third Defendant) where the defendants. The dispute stemmed from the then Prime Minister's announcement on government funding to expand the KickStart Breakfast Programme which aimed to provide free breakfast to children in lower decile schools.[290] The crux of the matter lay in the publication of cartoons by the Defendants that portrayed Māori and Pasifika communities in a negative and unfavourable manner.

The Plaintiff argued that the cartoons depicted people of Māori and Pasifika background as "lazy, neglectful, gluttonous, smokers and drinkers" and thereby were violating section 61 of the HRA. [291] Fairfax contested this, asserting that the case should be interpreted by striking a balance between the right of freedom of expression, as enshrined in the NZBORA, and determining the extent to which this right is limited by section 61 of the HRA.[292]

[289] *Wall v Fairfax New Zealand Ltd* [2018] NZHC 104.
[290] *Wall v Fairfax New Zealand Ltd* [2017] NZHRRT 17.
[291] *Wall v Fairfax New Zealand Ltd* [2018] NZHC 104 at [13] and Human Rights Law Centre "New Zealand High Court finds insulting cartoons did not breach hate speech legislation" (2018) < https://www.hrlc.org.au/human-rights-case-summaries/2018/6/1/new-zealand-high-court-finds-insulting-cartoons-did-not-breach-hate-speech-legislation>.
[292] *Wall v Fairfax New Zealand Ltd* [2017] NZHRRT 17 at [42].

In *Wall v Fairfax New Zealand Ltd*, the Tribunal and the High Court applied the two-limb test under s 61 of the Human Rights Act 1993.The first limb requires the expression in question to be "threatening, abusive, or insulting". The second limb asks whether that expression is "likely to excite hostility against or bring into contempt" a group of persons based on their colour, race, or ethnic or national origins. If both limbs were met, the question now was to determine which interpretation of section 61 of the HRA constituted the least possible limitation on the right of freedom of expression.[293] In *Wall v Fairfax New Zealand Ltd*, the cartoons were found to be insulting (satisfying the first limb). However, the courts determined that they did not satisfy the second limb, as they were unlikely to incite hostility or contempt in the broader public. The courts emphasised that section 61 sets a high threshold aimed at only the most egregious instances of racial vilification. This narrow construction of s 61 has prompted criticism for limiting its applicability in cases of subtle or systemic online hate, particularly in digital environments where repetition and virality can exacerbate harm.

The Tribunal applied an objective test to examine the matter and found the cartoons to be insulting. However, it was also necessary to establish whether a reasonable person would believe that such content was likely to incite hostility or bring contempt upon any group of individuals in New Zealand. In this regard, it was deemed unlikely to have such an effect. Subsequently, the plaintiff appealed the decision to the High Court.

Building on the previous discussion, the High Court unanimously found that the cartoons in question did not contravene section 61 of the Human Rights Act. In reaching this conclusion, the Court considered several important factors: First, the Court reaffirmed New Zealand's international obligations. It recognised that, as a member of the United Nations and a signatory

[293] *Wall v Fairfax New Zealand Ltd* [2017] NZHRRT 17 at [173].

to key international instruments such as the Universal Declaration of Human Rights and the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), New Zealand is bound by Article 4 of ICERD, which condemns propaganda and organisations that promote racial hatred or discrimination. The Court also referred to Article 19 of the International Covenant on Civil and Political Rights (ICCPR), which protects freedom of expression but allows lawful restrictions necessary for the respect of the rights or reputations of others.[294] Second, the Court stressed the need to apply the two-limb test established under section 61. This involves determining whether the words used were threatening, abusive, or insulting, and whether they were likely to incite hostility or bring a group into contempt. Finally, the Court emphasised the high threshold for racist speech under section 61. The prohibition applies only to relatively egregious forms of expression that inspire enmity, extreme ill-will, or are likely to result in a group being despised. This interpretation drew on dictionary definitions and comparative Canadian cases such as *Taylor v Canadian Human Rights Commission and Whatcott*, which defined "hatred" and "contempt" as expressions involving detestation, vilification, or the act of treating a group as inferior or unworthy.[295]

In applying this threshold, the Court noted that mere offence, humiliation or insult does not meet the legal standard. Instead, there must be real and substantial risk that the communication could provoke serious animosity or lead a reasonable person to adopt hostile or contemptuous views toward the targeted group. The Court explicitly rejected a broader interpretation that would allow claims based solely on "self-contempt" or personal distress within the group targeted, emphasising instead that the provision targets harmful intergroup dynamics, not subjective offence. The Court also reiterated that a reasonable person must assess the likelihood

[294] United Nations Human Rights *International Covenant on Civil and Political Rights* (23 March 1976).
[295] *Wall v Fairfax New Zealand Ltd* [2018] NZHC 104 at [53].

of harm occurring, considering the full context of the publication. For example, cartoons in satirical contexts would not typically meet the threshold unless they evoked deep, widespread, and objectively measurable hostility.

This interpretation aligns with Geddis' observation that New Zealand courts adopt a context-sensitive and deferential approach in freedom of expression cases. He notes that courts are cautious about overstepping into areas of contested public morality, and tend to interpret limitations on speech narrowly, especially where the statutory language is not explicit about the kind of harm intended to be captured.

This narrow judicial approach, though protective of expressive freedom, can risk underestimating the cumulative impact of racialised speech on marginalised communities. While it is essential to avoid criminalising mere offence, the current standard may be too rigid to address more systemic or covert forms of hate speech online. As this thesis argues, courts should consider the evolving nature of digital communication and reassess whether the threshold under section 61 still serves the Act's remedial purposes in today's online environment.

The Court further held that where multiple interpretations of a statute are possible, the one that least restricts the right should be preferred, provided such an interpretation is reasonably open.[296] In applying this principle to the censorship context, the Court held that terms such as "promotes or supports" in section 3 of the Film, Video, and Publications Classification Act 1993 must be interpreted narrowly. It ruled that the words should mean active encouragement

[296] *Moonen v Film And Literature Board Of Review* [2000] 2 NZLR 9 (CA) at [17].

or advocacy, not mere depiction or description.[297] This interpretive restraint sought to prevent vague or overbroad limits on expression.

Although *Moonen* concerned censorship law, the six evaluative criteria it set out (including the dominant effect of the publication, the character and medium of the expression, the target audience, and the publication's intended use) have informed broader approaches to balancing expression and harm.[298] The High Court in *Wall* referred to these to support a nuanced, contextual analysis of whether the cartoons crossed the threshold into prohibited hate speech. *Wall* confirms the very high hostility threshold under s 61: even racially stereotyped cartoons did not qualify. In practice, s 61 offers limited protection in digital contexts where repetition and virality amplify harm**.** This outcome sits uneasily with Waldron's dignity theory, which emphasises the assurance to minorities that they are accepted as equal citizens. Fredman's account of substantive equality likewise shows how high thresholds perpetuate systemic disrespect.

In my view, the *Moonen* framework is highly relevant to online hate speech regulation. In the context of social media, content is disseminated rapidly and often stripped of context, making interpretive caution even more essential. Courts must weigh not only the literal content of a post but also its virality, reach, tone, and intended impact. A static or overly narrow application of section 61 risks underestimating how digital communication amplifies harm. Therefore, a rights-consistent approach like that in *Moonen* should be applied dynamically to reflect the realities of digital expression.

[297] *Moonen v Film And Literature Board Of Review* [2000] 2 NZLR 9 (CA) at [29].
[298] The Classification Office - Te Mana Whakaatu "Christchurch Mosque Attack Livestream" (2019) <https://www.classificationoffice.govt.nz/news/featured-classification-decisions/christchurch-mosque-attack-livestream/>.

The decisions in *Moonen* and *Wall* illustrate the cautious and structured approach New Zealand courts adopt when balancing freedom of expression with the need to prevent harm. In *Moonen*, the Court of Appeal set out a proportionality framework for assessing whether restrictions on rights under the NZBORA are demonstrably justified. It emphasised that vague or broad statutory terms must be interpreted narrowly to avoid overreach, particularly in relation to expressive freedoms. This approach was echoed in *Wall*, where the High Court considered whether racially offensive cartoons published in mainstream newspapers breached section 61 of the Human Rights Act 1993. Despite acknowledging that the cartoons were objectively insulting, the Court concluded they did not reach the high threshold of inciting "extreme ill-will" or causing the group to be "despised." Both cases confirm that the courts maintain a high threshold for hate speech claims, and that mere offence or insult, even when widely perceived as racist, is insufficient. These rulings also demonstrate the courts' preference for a contextual interpretation of speech, taking into account tone, medium, and audience, which poses new challenges in the digital era where social media content is rapidly disseminated and difficult to contain. Importantly, the courts have shown a consistent reluctance to expand statutory provisions beyond their clear wording, deferring instead to Parliament to define the scope of harmful expression. In doing so, the judiciary preserves a wide margin for freedom of expression but may leave racialised communities vulnerable to more subtle and systemic forms of online hate speech. This thesis argues that such judicial restraint, while doctrinally consistent, highlights the need for legislative reform to ensure New Zealand's legal framework meaningfully addresses the realities of digital harm.

While section 61 provides a civil avenue for addressing hate speech, section 131 of the Human Rights Act introduces a criminal offence for inciting racial disharmony. However, the evidentiary and procedural hurdles are considerably higher, raising further questions about

whether New Zealand's legal framework offers meaningful protection against racially motivated speech. In practice, prosecutions under section 131 are almost non-existent. The requirement to prove intent beyond reasonable doubt, combined with the high expressive-freedom threshold under section 14 of NZBORA, makes it difficult for the Crown to secure convictions. The last attempted prosecution was *King-Ansell v Police*[299], decided under the earlier Race Relations Act, where the Court held that antisemitic statements were caught by the provision under section 25 of the Race Relations Act 1971. This case involved Colin King-Ansell, leader of the National Socialist Party of New Zealand, who was prosecuted for publishing an antisemitic pamphlet. The key legal issue was whether Jews in New Zealand could be considered a group with "ethnic origins" under the statute. [300]

The Court of Appeal unanimously held that they could. Richmond J and Woodhouse J emphasised that the term "ethnic origins" should be interpreted broadly in line with the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), which the Act implements.[301] In reaching that view, the Court engaged in a careful analysis of the statutory language and the Convention's purpose**.** Richmond J held that the phrase must be read broadly and in its popular sense, such that each term (race, ethnic, and national origins) "gives colour to the meaning of the others."[302] His Honour concluded that Jewish people in New Zealand form a group with shared customs, beliefs, and history, making them identifiable under the Act.

Woodhouse J affirmed this approach, emphasising that racial discrimination should be interpreted broadly to give effect to New Zealand's obligations under the International

[299] *King-Ansell v Police* [1979] 2 NZLR 531 (CA).
[300] *King - Ansell v Police* [1979] 2 NZLR 531 (CA) at [532] per Richmond J.
[301] *King - Ansell v Police* [1979] 2 NZLR 531 (CA) at [540] & [543].
[302] *King - Ansell v Police* [1979] 2 NZLR 531 (CA) at [532].

Convention on the Elimination of All Forms of Racial Discrimination (ICERD).[303] He warned against narrow or biological definitions of race and ethnicity, noting that shared social identity, belief systems, and cultural continuity were sufficient to meet the statutory standard.[304]

Richardson J similarly underscored the need for a contextual and purposive interpretation that is consistent with both the Convention and evolving dictionary definitions. His Honour stated that ethnic origins include "a sufficient combination of shared customs, beliefs, traditions and characteristics derived from a common or presumed common past".[305]

The Court also reaffirmed that criminal liability under section 131 (then section 25) hinges on demonstrable intention**,** that is, a purposive attempt to stir hostility or ill-will based on race or ethnicity.[306] The Court rejected arguments that antisemitism targeted religion rather than race, holding instead that Jewish people shared sufficiently distinct ethnic characteristics to fall under the statute. This decision clarified both the protected groups and the high threshold of intent required for conviction.

This case is critical to the thesis for several reasons. First, it affirms that hate speech law in New Zealand can recognise symbolic and cultural identity as a basis for protection, not just biological difference. Second, it illustrates how courts may balance freedom of expression with the need to protect minority communities from targeted hostility, by interpreting statutory language in light of international commitments and social context. From a contemporary perspective, the decision in *King-Ansell* suggests that section 131 could be activated in hate speech cases involving online abuse, particularly where content is deliberately aimed at

[303] *King - Ansell v Police* [1979] 2 NZLR 531 (CA) at [540] & [543].
[304] *King - Ansell v Police* [1979] 2 NZLR 531 (CA) at [534].
[305] *King - Ansell v Police* [1979] 2 NZLR 531 (CA) at [543].
[306] *King - Ansell v Police* [1979] 2 NZLR 531 (CA) at [541].

vilifying an ethnic group. However, its lack of use since the 1970s also reveals the inertia or hesitation in prosecuting under this provision, possibly due to the high threshold of intent, the requirement of Attorney-General consent (s 132), and a reluctance to curtail speech in borderline cases.[307]

King-Ansell recognised group-based dignity harms and set a robust understanding of "hostility" and "contempt," but its age and pamphlet-era context limit its guidance for online hate. Later courts have not extended it into a consistent hate speech jurisprudence. While King-Ansell resonates with Waldron's concern for protecting vulnerable groups from public hostility, its limited uptake suggests that dignity-based and equality-based approaches have not shaped subsequent New Zealand jurisprudence.

King-Ansell offers a clear legal foundation for recognising the collective harm of hate speech targeting ethnic identity but also exposes the inertia within New Zealand's criminal framework. Despite its doctrinal significance, section 131 has rarely been invoked since this case. This may reflect a broader reluctance to criminalise speech, evidentiary difficulties, or the procedural hurdle of obtaining Attorney-General consent. As this thesis explores, this reluctance may be increasingly unsustainable in the digital era, where online hate speech often spreads virally and anonymously, targeting marginalised communities in ways that civil remedies cannot effectively address. The narrow construction of section 61 and the rarity of prosecutions under section 131 have left significant gaps, particularly in the digital sphere. The Harmful Digital Communications Act 2015 was introduced to fill some of these gaps, but its focus on interpersonal harm rather than systemic hostility shows the continuing limits of New Zealand's

[307] Human Rights Act 1993, s 132 - stipulates: no prosecution of an offence against section 131 shall be instituted without the consent of the Attorney-General.

approach. The shortcomings of the Human Rights Act 1993 in addressing systemic or online hostility highlight the need for a more targeted regulatory response; one that acknowledges the speed, reach, and permanence of digital communication.

Since then, the criminal offence has been largely dormant. Scholars and the Human Rights Commission have described section 131 as a symbolic measure-important for signalling social condemnation of racial incitement, but ineffective as a practical deterrent. The rarity of prosecutions reinforces the broader concern that New Zealand's current framework provides limited recourse against serious hate-motivated expression. The high evidentiary burden and narrow scope of section 131 illustrate how the law prioritises expressive freedom over group protection, leaving significant gaps in enforcement, particularly for online content.

4.5 Harmful Digital Communications Act 2015 ("HDCA")

The HDCA was enacted following the Law Commission's 2012 conclusion that existing laws were inadequate for the "speed, reach, and permanence" of digital harm, providing a clear framework for responding to online abuse, cyberbullying, and image-based harassment.[308] It establishes both criminal and civil mechanisms for redress, including the ten Communication Principles in section 6 and the role of Netsafe as Approved Agency. The Law Society and media organisations warned that the proposed criminal offence in section 22 risked chilling satire or robust critique.[309] In response, the Justice and Electoral Committee narrowed the offence and clarified the intention requirement. [310]

[308] Law Commission Harmful Digital Communications: The Adequacy of the Current Sanctions and Remedies (Ministerial Briefing Paper, August 2012) at 2.4
[309] Law Commission *Harmful Digital Communications: The adequacy of the current sanctions and remedies* (Ministerial Briefing Paper, August 2012, Wellington) (Law Commission Paper).
[310] Justice and Electoral Committee *Harmful Digital Communications Bill: Report of the Committee* (March 2014) at 4

Since its enactment, scholars diverge on effectiveness. Geddis argues the Act filled a regulatory gap, but its procedural complexity and mediation-first model may deter complainants.[311] Other scholars contend the HDCA does not squarely address the structural roots of hate speech and online harm. Asafo argues New Zealand law tends to protect a "freedom of expression of racism," proposing a public-health reframing to capture cumulative impacts on Māori and Pasifika communities.[312] Panzic warns of unintended consequences for satire, whistleblowing, and political dissent.[313] Together, these critiques highlight the mismatch between an individual-harm model and group-based or systemic hostility.

Netsafe, designated under section 8, receives complaints, uses advice and mediation, liaises with platforms, and educates the public. However, it lacks coercive powers and cannot compel takedowns or unmask anonymous users.[314] This design choice was debated during passage of the Bill. Submitters questioned whether a purely advisory body could handle serious or persistent abuse; the Committee retained the mandate but recommended ongoing review. Survey evidence in 2018 suggested mixed satisfaction, especially where virality and anonymity limited outcomes.[315] The absence of a comprehensive, independent review since then leaves questions about effectiveness in cases involving sustained or hate-based abuse.

Section 6 sets out ten Communication Principles that guide assessment and remedies. For this thesis, Principles 8 and 10 are most relevant to hate-related content.[316] Principle 8 addresses

[311] Geddis, above n 262.
[312] Dylan Asafo "The Western Legal Roots of the Christchurch Mosque Shootings: A Colonial Critique of New Zealand's Legal Framework on Racist Hate Speech" (2021) 12 *UC Irvine L Rev* 101 at 17
[313] Stephanie Panzic "The Harmful Digital Communications Act 2015: A Deficiency in Defining Harm and an Unintended Limitation on Freedom of Expression" (2015) 21 *Auckland U L Rev* 218 at 243
[314] Harmful Digital Communications Act 2015 , s 8.
[315] Edgar Pacheco and Neil Melhuish *Harmful Digital Communications in New Zealand: Annual Population Survey 2017* (Netsafe, January 2018) at 38.
[316] Harmful Digital Communications Act 2015, s 6.

coordinated harassment by prohibiting incitement or encouragement to send messages to an individual for the purpose of causing harm.[317] Principle 10 states that digital communication should not denigrate an individual by reason of colour, race, ethnic or national origins, religion, gender, sexual orientation, or disability.[318] These provisions offer partial protection where hate is directed at an identifiable person, but they do not create a general prohibition on hate speech or enable group-based claims.

Section 22 creates a criminal offence where a person posts a digital communication with the intention to cause harm, the communication would cause harm to a reasonable person in the position of the target, and harm did occur.[319] "Harm" is defined in section 4 as serious emotional distress, a threshold designed to exclude trivial irritation while capturing significant emotional injury. This threshold reflects the Law Commission's view that emotional harm can be inferred from context and does not require clinical diagnosis.

In *Police v B,*[320] semi-undressed images of a complainant were posted on Facebook. The District Court dismissed the charge for lack of proof of serious emotional distress. On appeal, Downs J clarified that serious emotional distress is assessed contextually, does not require clinical diagnosis, and may be inferred from intensity, duration, manifestation, and context.[321] The appeal succeeded. The case confirms the HDCA's capacity to address non-trivial digital abuse while exposing evidential challenges where harm accrues cumulatively.

[317] Harmful Digital Communications Act 2015, s 6.
[318] Harmful Digital Communications Act 2015, s 6.
[319] Harmful Digital Communications Act 2015, s 22.
[320] *Police v B* [2017] NZHC 526.
[321] *Police v B* [2017] NZHC 526 at [24] and [43].

In *Hooper v Gee*, the High Court found grossly offensive and harassing posts caused serious emotional distress and affirmed that HDCA restrictions must be proportionate and consistent with section 14 of NZBORA.[322] The Court narrowed an overly broad District Court order and set a finite period of restraint.[323] Read with *Police v B*, the case law shows a maturing approach that recognises digital harm but calibrates remedies against expressive rights.

Empirical indicators suggest racialised online abuse persists. Netsafe's 2017 data showed disproportionately high rates of online abuse targeting Asian, Māori, and Pacific peoples, and the Human Rights Commission recorded substantial race-related complaints in 2018-19.[324] While dated, these figures align with accounts of continued racialised harm online and underscore the limits of relying on interpersonal frameworks where hostility is distributed, ambient, and platform-amplified.

The HDCA offers important tools for interpersonal digital abuse, including court-ordered remedies and a criminal offence tailored to serious emotional distress. It does not, however, directly address group-based or systemic hate speech, nor the amplification dynamics of social media. Within this chapter's wider argument, the HDCA should be understood as a necessary but partial response that complements, rather than replaces, targeted hate speech regulation.

In line with the chapter's argument, the HDCA was a necessary first step but remains ill-equipped to address group-level harms or algorithmically amplified hostility, making reform essential. The Act provides District Court remedies under section 19 (takedown and cease-and-desist orders, and in some cases compensation).

[322] *Hooper v Gee* [2022] NZHC 1854
[323] *Hooper v Gee* [2022] NZHC 1854 at [162]-[164]
[324] Boundy, above n 7.

The Act was introduced aimed to provide a clear legal framework for responding to a wide range of online harms, including cyberbullying, harassment, and the unauthorised dissemination of personal or intimate content. It aims to protect individuals from such damage and to promote accountability by regulating digital conduct that crosses certain thresholds. Under the HDCA, harmful digital communication refers to a message or series of messages that are targeted at an individual, cause serious emotional distress, and breach one or more of the ten Communication Principles set out in section 6 of the Act. These principles define unacceptable online behaviours, such as threatening messages, disclosures of sensitive personal information, or attacks based on race, gender, or sexual orientation. The threshold of serious emotional distress has been interpreted by the courts and is explored further below in the discussion of *Police v B*.[325]

The HDCA offers complainants a set of remedies issued by the District Court under section 19 of the Act, including takedown orders, cease-and-desist orders, and, in some cases, compensation for harm suffered.[326] These judicial mechanisms aim to provide individuals with accessible and timely recourse when they have been subjected to harmful online conduct.

These principles guide the assessment of complaints and are central to the Act's purpose: to "deter, prevent and mitigate harm caused by digital communications and to provide victims of harmful digital communications with a quick and efficient means of redress"[327]. However, while the HDCA provides protection from a range of online harms, including some forms of discriminatory or abusive speech, it does not explicitly regulate hate speech as a distinct legal

[325] *Police v B* [2017] NZHC 526.
[326] Harmful Digital Communications Act 2015, s 19.
[327] Edgar Pacheco and Neil Melhuish "Annual Population Survey 2017" (2018) Netsafe - Online Safety Help and Advice for New Zealanders <https://www.netsafe.org.nz/annual-population-survey-2017/>.

category. For example, there is no standalone offence of inciting hatred based on religion, gender identity, or sexual orientation, and the Act is framed around individual harm, rather than the protection of targeted communities or social groups. In this sense, New Zealand lacks comprehensive legislation that addresses group-based hate speech, especially where harm is systemic or ideological rather than personal.

This distinction is important for the purposes of this thesis: although the HDCA provides a useful framework for managing interpersonal digital harm, it does not fully address the broader societal and democratic risks posed by hate speech online.

To be clear, Principles 8 and 10 operate only where an identifiable individual is targeted; they do not create a general prohibition on hate speech or enable group-based claims. This limitation is central to the thesis's account of fragmentation.

According to the HDCA, it is a criminal offence if a person uploads or posts digital communication with the intention to cause harm.[328] The offence requires proof of intention to cause harm, objective foreseeability of harm to a reasonable person in the target's position, and evidence that harm occurred. The term "harm" is defined as "serious emotional distress".[329] The Law Commission's analysis influenced this threshold and supports a blended objective-subjective assessment.

[328] Harmful Digital Communications Act 2015, s 22.
[329] Harmful Digital Communications Act 2015, s 22(1).

In *Police v B,*[330] Downs J confirmed that serious emotional distress is a value-laden factual assessment that may be inferred from the totality of circumstances.[331] It also exposes evidential gaps where harm is cumulative, ambient, or symbolic.

Another significant decision is *Hooper v Gee*, where the High Court both confirmed the serious-emotional-distress standard and emphasised NZBORA-consistent, proportionate orders.[332] In this case, the respondent sought a takedown order under section 19 after the appellant posted defamatory and inflammatory material on Facebook, falsely linking the respondent to child exploitation.[333] The District Court granted a broad, indefinite restriction. On appeal, the High Court found that the communication was grossly offensive, harassing, and caused serious emotional distress; therefore, satisfying the test under section 22(1) HDCA and the applicable Communication Principles.[334]

The Court also undertook a NZBORA analysis, affirming that while the HDCA limits freedom of expression, those limits are justifiable under section 5 of NZBORA. Relying on legislative history, Crown Law advice, and sections 6 and 19 of the HDCA itself, the Court concluded that restrictions under the Act must be interpreted as proportionate and rights-consistent.[335] Consequently, Fitzgerald J replaced the District Court's open-ended prohibition with a more narrowly tailored order, restricting the appellant from posting about the respondent until 1 August 2023.[336]

[330] *Police v B* [2017] NZHC 526.
[331] *Police v B* [2017] NZHC 526 at [43]
[332] *Hooper v Gee* [2022] NZHC 1854
[333] *Hooper v Gee* [2022] NZHC 1854 at [2] and [50]
[334] *Hooper v Gee* [2022] NZHC 1854 at [162]-[164]
[335] *Hooper v Gee* [2022] NZHC 1854 at [104]-[109]
[336] *Hooper v Gee* [2022] NZHC 1854 at [173]

This judgment complements *Police v B* by confirming that "serious emotional distress" does not require clinical diagnosis but must be assessed holistically. It also underscores the courts' obligation to interpret HDCA remedies in a manner consistent with expressive rights. Together, these cases reflect a maturing body of jurisprudence that both recognises digital harm and preserves constitutional freedoms. Taken together, *Police v B* and *Hooper v Gee* illustrate the evolving judicial interpretation of "harm" and the careful calibration courts must apply when navigating between victim protection and freedom of expression.

Data collected by Netsafe in 2017 indicated disproportionately high rates of online abuse targeting Asian, Māori, and Pacific individuals, highlighting the racialised nature of digital harm.[337] While this data predates the Christchurch mosque attacks in 2019, more recent reports suggest that online hate remains persistent and in some cases, intensified. For example, the New Zealand Human Rights Commission received 1,417 complaints of unlawful discrimination in the 2018/19 reporting period, of which 369 related to race, colour, or national or ethnic origin. [338] These figures highlight the enduring prevalence of race-based harm in both physical and digital contexts. While the data predates the Christchurch mosque attacks and more recent global shifts such as the COVID-19 pandemic and social justice movements, it underscores the systemic nature of racial abuse and the limitations of existing redress mechanisms. This trend supports the need for legal tools like the HDCA to respond to digital communications that cause emotional harm, particularly where such communications intersect with race, identity, and structural inequality. The *Police v B* decision thus sits within a broader context of increasing demand for legal frameworks that can account for the lived experiences of victims of racialised digital abuse.

[337] Boundy, above n 7.
[338] New Zealand Human Rights Commission Annual Report 2018/19 (2019) at 36.

While the HDCA offers important redress for victims of digital abuse, especially where serious emotional distress is proven, its scope remains primarily interpersonal. It does not directly address the collective and structural harms of hate speech, which erode equality and social cohesion beyond individual distress. The *Police v B* decision illustrates how courts interpret the Act conservatively, while the data on racialised online abuse shows that systemic harms remain widespread. Together, the limits reveal that the HDCA is better understood as a supplement to, rather than a substitute for, dedicated hate speech regulation. The continuing reliance on civil harm-based remedies underscores the inadequacy of New Zealand's broader statutory framework, a gap further highlighted when the Crimes Act and related legislation are considered in the following section. As shown in Chapter 3, algorithmic amplification compounds these risks, yet the HDCA does not engage with platform architecture. The next section considers how the Crimes Act and related legislation further illustrate these systemic gaps.

4.6 Other Statutes

New Zealand law contains several additional provisions that touch on harmful or offensive expression, but none directly establish a coherent framework for hate speech. Instead, these statutes regulate specific contexts, such as disorderly conduct in public places, broadcasting standards, or objectionable publications, through piecemeal rules. While each legislation has some relevance to online hate speech, their thresholds are either too narrow (capturing only situational disorder or imminent threats) or too broad and reactive (focused on extreme cases like violent imagery). This section examines the Summary Offences Act 1981, the Crimes Act 1961, the Broadcasting Act 1989, and the Films, Videos, and Publications Classification Act

1993, showing how each contributes fragments of regulation but leaves the wider problem unresolved.

4.6.1 Films, Videos, and Publications Classifications Act 1993 ("The Classifications Act")

Although the Classification Act is primarily designed to restrict extreme content such as child exploitation or violent publications, it has some limited relevance to hate speech. In particular, section 3(3)(e) addresses material that portrays groups as inherently inferior. However, the statute sets a high threshold and is reactive rather than preventative, making it a blunt tool for digital hate speech.

The Films, Videos, and Publications Classification Act 1993 defines a publication as "objectionable" if its availability is likely to be injurious to the public good. [339] In practice, the Act has been used mainly for extreme content such as sexual exploitation, torture, or violence. [340] It also extends to publications that promote the inferiority of groups on protected grounds under the Human Rights Act 1993, such as race, religion, disability, or sexual orientation.[341] This creates a possible, though rarely used, pathway for restricting hate material, including extremist symbols or Holocaust denial. From a theoretical perspective, the Act reflects Mill's harm principle by focusing on extreme injury before intervention. Waldron's dignity theory and Fredman's substantive equality approach suggest this leaves structural or coded hate outside its reach. Lessig's "architecture" lens also shows the problem, the Act does not account for how digital platforms amplify objectionable content before classification can occur.

[339] Film, Video, and Publications Classifications Act 1993, s 3(1).
[340] Film, Video, and Publications Classifications Act 1993, s 3(2).
[341] Film, Video, and Publications Classifications Act 1993, s 3(3)(e).

The Classification Act also includes controls on offensive language. Section 3A allows a publication to be age-restricted if it contains highly offensive language, where unrestricted access would be likely to cause serious harm to minors.[342] The test is framed around risk to children, not whether the words are harmful to society in a broader sense.[343] Responsibility for administering and enforcing the Act lies with the Office of Film and Literature Classification, led by the Chief Censor, and the Film and Literature Board of Review. The Chief Censor's Office classifies films, games, and publications referred by the public, the police, or the Department of Internal Affairs (DIA). The DIA has primary enforcement powers under the Act, including investigating and prosecuting offences involving objectionable material, such as extremist or hate-based publications distributed online. Material deemed "objectionable" is banned outright, while content that is age-restricted (for instance under section 3A) may only be supplied or viewed subject to classification labels. This narrow focus reflects a protective rather than structural lens. It limits the provision's ability to address discriminatory or hate-based language, which may not always cause immediate harm to children but still undermines equality and social cohesion in the long term.

This raises the question of what the legal test for "likely to cause serious harm" actually means. Harvey notes that the Classification Office applies a contextual test under section 3(4), weighing the dominant effect of the publication, the language used, the audience, and any literary or educational value.[344] The Court of Appeal in *Moonen v Film and Literature Board of Review* confirmed that classification requires balancing the public good against the right to freedom of expression in section 14 of NZBORA.[345] The Court stressed proportionality and

[342] Film, Video, and Publications Classifications Act 1993, s 3A.
[343] Film, Video, and Publications Classifications Act 1993, s 3A.
[344] David Harvey *internet.law.nz* (Fifth ed, LexisNexis NZ Limited, 2023), at 203.
[345] *Moonen v Film and Literature Board of Review* [2000] 2 NZLR 9 (CA) at [16]

held that where the harm is marginal, freedom of expression must prevail. Harvey comments that this shows how contextualised restrictions are favoured over absolute bans.[346] Yet this standard, while rights-protective, sets a high threshold and makes it difficult to restrict coded or systemic hate speech that falls short of obvious serious harm.

The Office has also applied this precautionary approach to content that promotes or supports harmful conduct, even where intent is absent. For example, an image taken by a minor of herself in a sexualised pose was deemed objectionable because of its potential to normalise exploitation, despite no third-party perpetrator being involved.[347] This emphasis on effect over intent, affirmed in *Moonen*, shows how the classification framework can capture indirect harms. However, the same logic has rarely been extended to ideologically coded hate material, which often evades regulation despite its corrosive social effects.

The Act provides a legally defined, multi-factor test for determining when a publication (including one containing hate-based or discriminatory messaging) may be restricted or deemed objectionable. However, in practice, this test is most effective against extreme or violent material. Its reliance on "serious harm to the public good" means that lower-level or coded hate speech is unlikely to meet the threshold, even when it causes significant harm for targeted communities online. This shows the Act's limits: it provides a strong tool against extreme cases but fails to address the more common, systemic forms of digital hate speech that erode dignity and equality over time.

[346] David Harvey *internet.law.nz* (Fifth ed, LexisNexis NZ Limited, 2023), at 202.
[347] David Harvey *internet.law.nz* (Fifth ed, LexisNexis NZ Limited, 2023), at 209.

Courts do not directly determine whether content is objectionable; this responsibility falls under the purview of the Office of Film and Literature Classification ("the Classification Office).[348] The Classification Office, headed by the Chief Censor, reviews films, games, and publications referred by the public, the police, or the Department of Internal Affairs. It applies the statutory test in section 3 of the Classification Act, which defines a publication as "objectionable" if it describes, depicts, or otherwise deals with matters such as sex, horror, crime, cruelty, or violence in a way that is likely to cause injury to the public good. The Office may classify material as unrestricted, restricted to certain age groups, or outright banned.[349] Its decisions can be reviewed by the Film and Literature Board of Review and, on questions of law, appealed to the High Court. This system means that the courts become involved only at the appellate stage, rather than making the initial determination themselves.

Section 41 of the Classification Act makes the Office's decision final and binding: once material is classified as objectionable, this is conclusive evidence in court. [350] Appeals are only possible on narrow points of law, such as whether the statutory definition of "objectionable" was applied correctly.[351] This gives certainty but also rigidity. It means courts cannot revisit the substance of the classification, even if questions arise about context or evolving social harm. For digital hate speech, where meaning is often coded or contested, such rigidity risks leaving harmful content outside the law's reach.

[348] Film, Video, and Publications Classifications Act 1993, s.77.
[349] Film, Video, and Publications Classifications Act 1993, s.29(c) Where in any civil or criminal proceedings the defendant admits that a publication—
(a) is objectionable; or
(b) is objectionable except in any 1 or more of the circumstances referred to in subsection (1)(b
[350] Film, Video, and Publications Classifications Act 1993, s41.
[351] Film, Video, and Publications Classifications Act 1993, s3.

Enforcement is carried out by the Department of Internal Affairs' Censorship Compliance Unit, which monitors online spaces and uses digital tools under the Search and Surveillance Act 2012. [352] In *Hutton v R*[353], covert DIA investigators used a digital "beacon" to unmask a dark web user distributing child exploitation material.[354] The Court of Appeal accepted the method as proportionate, recognising the need for investigative flexibility in online harm. [355] While this case involved child exploitation, the framework could theoretically extend to hate material in covert forums. Yet its focus on extreme abuse again shows the system's limits: tools are available, but they have not been applied to structural or symbolic hate speech.

Under the Search and Surveillance Act 2012, the Classification Office's determination that material is "objectionable" provides the legal basis for enforcement action. [356] Once content is classified as objectionable, warrants may be issued to seize relevant digital material. [357] While these powers equip authorities to respond to extreme content, their focus remains on child exploitation or violent imagery, not systemic or coded hate speech.

These provisions give law enforcement and regulatory agencies the legal authority and procedural mechanisms to investigate, identify, and respond to objectionable content and, in some circumstances, online hate speech. This responsibility is primarily carried out by the Censorship Compliance Unit, a division within the Department of Internal Affairs (DIA), and the Digital Child Exploitation Team (DCET), which forms part of the DIA's Digital Safety Directorate. The DCET conducts specialist investigations into online abuse and works in collaboration with international partners, including INTERPOL and Five Eyes intelligence-

[352] Search and Surveillance Act 2012, s.71.
[353] *Hutton v R* [2018] NZCA 419.
[354] *Hutton v R* [2018] NZCA 419 at [9].
[355] *Hutton v R* [2018] NZCA 419 at [49].
[356] Film, Video, and Publications Classifications Act 1993, s.109.
[357] Film, Video, and Publications Classifications Act 1993, s.109A.

sharing networks, to trace, monitor, and disrupt networks engaged in serious digital harm, including child sexual exploitation and other criminal publications. These teams are equipped to investigate extreme digital harms such as child exploitation and sometimes objectionable publications. Online hate speech, unless it clearly falls within "objectionable" material, rarely triggers equivalent scrutiny.

Regarding visual content, the Classification Act can restrict material under section 3(3)(e) if it portrays a group as inherently inferior on prohibited grounds such as race, religion, disability, or sexual orientation.[358] In principle, this could apply to extremist symbols or manipulated images that promote group inferiority. However, the threshold remains high: unless the content explicitly advocates hatred or discrimination, many forms of coded or symbolic hate are unlikely to be captured. This highlights the gap between the Act's formal scope and its limited practical impact on digital hate imagery.

A case study that demonstrates both the reach and the limits of the Act is the Christchurch mosque attacks. Following the livestream of the attack in 2019, the Classification Office declared the footage "objectionable" under s 3.[359] While effective in banning the immediate content, it also exposed the Act's limits: rooted in 1993 concepts like "injury to the public good," the framework struggles with contemporary digital ecosystems.[360] Coded references (such as memes, numbers, or slogans) may be legible to extremist audiences but remain outside traditional harm tests, creating gaps in coverage.[361]

[358] Film, Video, and Publications Classifications Act 1993.
[359] The Classification Office - Te Mana Whakaatu "Christchurch Mosque Attack Livestream" (2019) <https://www.classificationoffice.govt.nz/news/featured-classification-decisions/christchurch-mosque-attack-livestream/>.
[360] The Classification Office - Te Mana Whakaatu, above n 356.
[361] The Classification Office - Te Mana Whakaatu, above n356.

Parliament responded with the 2022 Urgent Interim Classification Amendment, which created livestream-specific offences and interim classification powers.[362] These changes improved responsiveness but remained crisis-driven, addressing only extreme cases while avoiding broader duties on platforms. More ambitious proposals, such as web filtering, were abandoned over freedom of expression concerns.[363]

The Christchurch livestream illustrates both the Act's utility and its limits. It shows how the statute can react to violent extremist content, but the high "injury to the public good" threshold leaves systemic and coded hate largely untouched. This gap reflects the dominance of Mill's harm principle in New Zealand's regulatory approach, privileging liberty until tangible injury occurs. On the other hand, as Waldron argues, hate speech erodes dignity and equal standing long before violence. Fredman's substantive equality framework similarly suggests that ignoring persistent online vilification entrenches disadvantage for minority communities. From Lessig's perspective, the Act's focus on content misses the role of "architecture," as platform design accelerates dissemination before classification can occur.

When considered together, these perspectives show why the current regime is reactive, piecemeal, and poorly adapted to the dynamics of digital hate speech. The Act's focus on objectionable publications reflects a media-era model. Yet, as discussed in Chapter 3, online harm often arises less from isolated content than from its amplification through digital architecture; this is a dimension the Act does not capture.

[362] Films Videos and Publications Classification Urgent Interim Classification of Publications and Prevention of Online Harm Amendment Bill (26 May 2020) 268-3.

[363] Currently, there is only one web filter that is government-backed that is developed to block sexual exploitation material (Digital Child Exploitation Filtering System). The Bill facilitates the potential future requirement to develop a web filter. Note, that this bill is only applicable to web page filtering and not messaging applications and other online services. The rationale is so that it does not impact the freedom of expression. The government's cautious tone around expanding filtering systems reflects a lingering tension between online safety and freedom of expression. However, its selective scope that is limited to child exploitation, reveals a reluctance to extend similar protections to hate-based content.

4.6.2 Summary Offence 1981

Although the Summary Offences Act 1981 regulates disorderly and offensive behaviour in public, its relevance to hate speech is marginal. Courts have interpreted "offensive" narrowly, applying it mainly to situational outbursts rather than systemic vilification. The Act highlights how older statutes struggle to address digital communication, where "public place" takes new forms online. The Summary Offences Act 1981 prohibits disorderly behaviour (s 3) and offensive language or conduct in public (s 4(1)), punishable by fines up to NZD $1,000.[364] On its face, this could capture threatening or insulting speech directed at minority groups in public settings. However, the provisions are framed around situational disruption and public order, not the dignitary or systemic harms of hate speech. Their relevance to online vilification is therefore limited from the outset.

The case of *Evans v R,* the appellant was convicted under s 4(1) after directing sexually explicit and inflammatory remarks at police officers during a large public concert.[365] The Court held that his words amounted to "insulting language" in a public place, rejecting arguments that the venue was not public and that the speech was protected by NZBORA.[366]

The reasoning, drawing on *Brooker v Police*[367] and referencing *Coleman v Power*[368] demonstrates how the Act targets situational disruption and offensive insults rather than systemic hostility.[369]

[364] Summary Offences Act 1981, s.4(1).
[365] *Evans v R* [2008] DCR 199.
[366] *Evans v R* [2008] DCR 199, at [209]
[367] *Brooker v Police* [2007] NZSC 30.
[368] *Coleman v Power* (2004) 220 CLR 1.
[369] *Evans v R* [2008] DCR 199 at [206]

Comparative jurisprudence increasingly recognises social media as a functional public spaces. In *Packingham v North Carolina*, the U.S. Supreme Court identified social media as the "modern public square", a venue essential for public discourse.[370] By contrast, New Zealand courts have not yet considered whether "public place" in the Summary Offences Act extends to digital environments. This gap means that much online hate speech falls outside the statute's reach, despite occurring in forums that now serve the same public functions as traditional physical spaces.

Scholars have argued for a functionalist understanding of "publicness" that reflects the realities of digital communication. Suzor critiques the public/private binary, advocating a focus on democratic engagement and user visibility.[371] Barendt similarly suggests rethinking public discourse, though his privileging of political over casual or commercial speech becomes difficult to sustain as platforms increasingly serve as de facto public spaces.[372] These perspectives strengthen the case for interpreting "public place" in the Summary Offences Act to include digital forums, yet New Zealand law has not taken this step.

This type of interpretation is important. If courts and lawmakers keep a narrow reading, the Summary Offences Act becomes outdated and cannot answer the challenges of harmful digital speech. A change, either by Parliament or by judicial interpretation, could let section 4(1) work better against online hate speech, while still protecting freedom of expression under NZBORA. But even with such changes, this Act is not a strong tool for hate speech. Courts usually give "offensive" a narrow meaning and prefer to protect expression when there is doubt. The law is

[370] *Packingham v North Carolina* 137 S. Ct. 1730 (2017) at [10]
[371] Nicolas Suzor, *Lawless: The Secret Rules that Govern Our Digital Lives* (Cambridge University Press, 2019) at 113.
[372] Robert C Post, 'Review of Freedom of Speech by Eric Barendt' (1988) 36(1) Am J Comp L 174, at 177

built to stop immediate disorder in public places, not long-term or systemic vilification. Because of this, the Act offers little real protection for marginalised groups who face the deeper harms of hate speech.

4.6.3 Crimes Act 1961

The Crimes Act 1961 provides offences for threats, incitement, and party liability, which in theory could capture certain extreme forms of online hate. However, these provisions presuppose an underlying offence and focus on physical or imminent harm rather than symbolic or cumulative hostility. The Act is therefore more relevant to violent extremism than to the broader spectrum of online hate speech.

Section 7 of the Crimes Act 1961 gives New Zealand jurisdiction over offences committed abroad if their effects are felt domestically.[373] In theory, this covers online hate speech where material is posted overseas but targets New Zealanders. However, jurisdiction alone does not create an offence: without a substantive hate speech prohibition, section 7 merely establishes territorial reach. It ensures courts could hear such cases, but it leaves a gap as there is no dedicated offence to prosecute.

Section 66 of the Crimes Act 1961 extends liability to those who aid, abet, incite, or procure another person's offence.[374] In principle, this could apply to individuals who coordinate or encourage online hate campaigns. Yet the provision presupposes an underlying offence. Because New Zealand lacks a stand-alone hate speech crime, s 66 can only attach liability

[373] Crimes Act 1961, s.7.
[374] Crimes Act 1961, s. 66(1)

where speech crosses into an existing criminal category (e.g. threats or harassment). This limits its relevance to the broader problem of systemic or ideological hate speech.

Section 66(2) creates liability where two or more people pursue a common unlawful purpose and an offence occurs as a probable consequence.[375] Courts have applied it in cases of group violence (for example in the case of *Ahsin v R*)[376], extending responsibility to foreseeable acts arising from collective intent. In theory, this could apply to coordinated digital hate campaigns, where participants share an unlawful aim and escalate harm together. However, without a substantive hate speech offence, it is difficult to characterise online hostility as an "unlawful purpose." The doctrine is therefore more illustrative than practical for hate speech: it exposes how criminal law has tools for collective wrongdoing, but they are tethered to physical violence rather than symbolic or cumulative online harms.

Still, translating these doctrines to platform-level responsibility is more speculative. Unlike users who incite or coordinate hate speech, social media companies are unlikely to be cast as "parties" unless there is strong evidence of intentional design or policy choices facilitating harm.[377] The Crimes Act's doctrines of aiding, abetting, or common purpose remain focused on individual culpability and interpersonal complicity. They are too narrowly drawn to regulate systemic or algorithmic facilitation by platforms. This thesis argues that while s 66 offers valuable tools for user-level or group-based digital hate, it falls short against institutional actors, underscoring the need for legislative reform or dedicated digital complicity offences.

[375] Crimes Act 1961, s. 66(2).
[376] *Ahsin v R* [2014] NZSC 153 at [102].
[377] Crimes Act 1961, s.66 (1) (d).

Building on the previous discussion of the Crimes Act's scope, section 69 of the Crimes Act 1961 imposes liability on individuals in New Zealand who aid, incite, counsel, or procure offences committed abroad, provided the act is also unlawful in the foreign jurisdiction. [378] In effect, this section extends criminal responsibility beyond national borders, but it is aimed primarily at serious extraterritorial crimes such as treason or espionage, not speech. Although it could, in theory, apply to cross-border digital incitement, the provision has never been used in this way. Its legislative design and history exclude preparatory or speech-based conduct, so in practice it remains peripheral to the regulation of online hate speech.

Flowing from the above, section 307A of the Crimes Act 1961 criminalises threats likely to endanger life, health, or safety.[379] In theory, its open-ended wording could capture certain forms of online hate speech involving threats. In practice, however, the courts interpret it narrowly. In *O'Neill v Malcouronne*, the High Court rejected an argument that a Facebook "friend request" breached the section, calling the claim "trivial" and outside the provision's scope.[380] The case confirms that s 307A is aimed at tangible threats to public safety or infrastructure, not the dignitary or psychological harms often produced by online hate. As a result, its relevance for hate speech is marginal at best.

The courts' narrow interpretation of s 307A keeps it tied to tangible threats of violence or public safety, not the symbolic or cumulative harms of hate speech. This means that online hostility, unless coupled with a credible threat of serious violence or disruption, falls outside

[378] Crimes Act 1961, s.69 (1): Party to any other crime outside New Zealand
Every one is liable to imprisonment for a term not exceeding 14 years who, in New Zealand, aids, incites, counsels, or procures the doing or omission outside New Zealand, by any person not owing allegiance to the Sovereign in right of New Zealand, of any act which, if done or omitted outside New Zealand by a person owing such allegiance, would be any of the crimes of treason, inciting to mutiny, or espionage, as specified in sections 73, 77, and 78.
[379] Crimes Act 1961, s.307A.
[380] *O'Neill v Malcouronne* [2021] NZHC 3027 at [4].

its ambit. The result is a doctrinal rigidity that exposes a wider weakness in New Zealand's criminal law: it lacks tools for addressing group-based or ideological harms that are socially corrosive but not easily classified as threats.

Section 311(2) of the Crimes Act 1961 extends criminal liability to individuals who encourage, counsel, or attempt to procure the commission of an offence, even if the incited offence is never carried out.[381] Judicial interpretation in *The King v Barker*, held that incitement becomes criminal when it is "of such a nature as to be in itself sufficient evidence of the criminal intent with which it is done," and that liability requires an overt act connected with the proposed crime.[382] This doctrinal focus on manifest intent means the provision can capture online speech that explicitly encourages violence or criminal acts. However, its relevance to hate speech is limited. Generalised hostility or ideological abuse, for example, calls to marginalise or vilify a group, falls outside the provision unless tied to a specific offence. As a result, s 311(2) remains a narrow tool: effective against explicit exhortations to crime, but unable to address the wider social and cumulative harms of online hate speech.

From a doctrinal perspective, these principles offer a potential basis for regulating certain forms of online hate speech, particularly where the content involves explicit encouragement to commit a criminal act, such as physical assault or property damage. However, the scope of section 311(2) is limited. It does not criminalise generalised hostility or ideological hate speech unless a specific offence is clearly encouraged. For example, urging others to "harass" or "get rid of" a particular ethnic group may fall outside the provision if it lacks a direct connection to

[381] Crimes Act 1961, s.311(2) Every one who incites, counsels, or attempts to procure any person to commit any offence, when that offence is not in fact committed, is liable to the same punishment as if he or she had attempted to commit that offence, unless in respect of any such case a punishment is otherwise expressly provided by this Act or by some other enactment.
[382] *The King v Barker* [1924] NZLR 865 at [875].

a defined criminal act. As such, while section 311(2) provides a prosecutorial route in extreme cases involving explicit incitement, it cannot address the broader phenomenon of structural or community-level hate speech. Its focus on individual intent and discrete acts illustrate the narrow reach of New Zealand's criminal law, leaving systemic harms and cumulative hostility beyond its scope.

Taken together, the HDCA and the Crimes Act represent two different approaches to regulating harmful expression. In contrast to the Crimes Act 1961, which relies on intent-based doctrines such as aiding, abetting, incitement, and threatening conduct, the HDCA adopts a harm-based model of civil redress. As explored earlier in Section 4.5, it provides important protections against digital abuse where serious emotional distress is involved. However, its focus on interpersonal harm, rather than group-based hostility, and its reliance on informal mechanisms such as mediation through Netsafe, limit its utility for addressing systemic or ideologically motivated hate speech.

While the HDCA fills some of the gaps left by the Crimes Act, particularly through Communication Principles 8 and 10, it is not designed to address hate speech as a public or democratic harm. The following section examines the Broadcasting Act 1989, which governs standards for television and radio content in New Zealand. This Act shifts the focus from criminal offences to content regulation, emphasising public accountability, fairness, and social responsibility rather than punishment.

4.6.4 Broadcasting Act 1989

The Broadcasting Act 1989 establishes standards against offensive and discriminatory broadcasting, but its jurisdiction is limited to traditional radio and television. With user-generated and on-demand content excluded, most digital hate speech falls outside its scope. This makes the Act an example of a regulatory gap: effective in its own sphere, but irrelevant to the online platforms where hate now spreads.

The Broadcasting Act 1989 establishes standards for radio and television content, overseen by the Broadcasting Standards Authority (BSA). Among its functions is the enforcement of the "discrimination and denigration" standard, which prohibits programmes that encourage the denigration of, or discrimination against, sections of the community on prohibited grounds such as race, sex, disability, or political and religious belief.[383] At first glance, this resembles a hate speech provision, as it seeks to safeguard community groups from harmful or stigmatising expression.[384]

However, the scope of the Act is tightly confined to traditional broadcast media. The BSA has confirmed that it does not regulate social media or user-generated online content, which now constitute the dominant spaces for the spread of hate speech.[385] Even within broadcasting, the Authority has applied the discrimination and denigration standard cautiously, intervening only where there is sustained malice rather than where content is merely offensive or satirical.[386] The BSA's 2024-2025 Statement of Performance Expectations reinforces this approach, emphasising balance between protecting audiences and upholding freedom of expression.[387]

[383] Broadcasting Act 1989, s. 21(1) (a)-(c).
[384] Broadcasting Standards Authority (2020) <https://www.bsa.govt.nz/about-us/what-we-do/>.
[385] Broadcasting Standards Authority "Does the BSA oversee content on social media"<https://www.bsa.govt.nz/all-faqs/online/ >.
[386] Broadcasting Standards Authority (2020) <https://www.bsa.govt.nz/about-us/what-we-do/>.
[387] Broadcasting Standards Authority, Statement of Performance Expectations 2024-2025 (June 2024) at 3

This legal conclusion is supported by the statutory definition of "broadcasting" in section 2(1), which explicitly excludes transmissions made solely in response to individual user requests. As social media and streaming services allow users to access content on demand, rather than receiving it simultaneously as in traditional broadcasting, they fall outside the scope of the Act. The BSA itself has acknowledged this limitation in its jurisdictional guidance.[388]

As such, while the Broadcasting Act contributes to New Zealand's wider patchwork of speech regulation, its relevance to hate speech is marginal. Its focus on traditional broadcasting leaves significant gaps in the digital environment, and its high interpretive threshold means that systemic or coded hostility is rarely addressed. This reinforces the argument advanced in section 4.2: New Zealand has institutionalised inadequacy by relying on fragmented statutory tools that are poorly adapted to contemporary forms of online hate. This narrow approach reflects Mill's liberty-focused model, where expression is protected unless it causes demonstrable harm, but fails to give effect to Waldron's concern with dignity and equal standing, leaving systemic hostility in broadcasting largely unaddressed.

4.7 Balancing Expression vs Regulation

The central tension in New Zealand's framework is the balance between freedom of expression under section 14 of the New Zealand Bill of Rights Act 1990 (NZBORA) and statutory restrictions on hate speech. Section 5 allows rights to be limited where "demonstrably justified in a free and democratic society."[389] The courts have applied this proportionality framework most clearly in *Moonen v Film and Literature Board of Review and R v Hansen* where the Court of Appeal and Supreme Court developed a structured approach for weighing rights

[388] Broadcasting Standards Authority (2020) <https://www.bsa.govt.nz/about-us/what-we-do/>.
[389] New Zealand Bill of Rights Act 1990, s 5.

against legislative objectives.[390] Yet in practice, proportionality analysis in the hate speech context has been limited. Courts have tended to give greater weight to expressive freedom than to the protection of vulnerable groups, resulting in high thresholds for liability.

This is evident in cases under s 61 of the Human Rights Act 1993, such as *Wall v Fairfax*. There, cartoons that portrayed Māori and Pasifika as "lazy" and "gluttonous" were deemed insulting but not sufficiently likely to "excite hostility or contempt.[391] Such reasoning privileges the liberty to insult over the assurance of dignity for minority groups. Waldron argues that equal citizenship requires not only freedom to speak but also freedom from being systematically portrayed as outsiders.[392] Similarly, Fredman's account of substantive equality highlights that laws which fail to address group vulnerability reproduce inequality under the guise of neutrality.

The political debate following the Christchurch Mosque attacks demonstrates how this judicial caution is mirrored in Parliament. Justice Minister Andrew Little argued that extending hate speech protections was necessary to prevent further social division, while ACT leader David Seymour framed such reforms as a threat to "free speech" and democratic debate.[393] This polarisation has produced legislative inertia: Labour withdrew its 2022 Religious Belief Amendment Bill due to lack of consensus, while ACT introduced its "Freedom to Speak Bill"

[390] *Moonen v Film and Literature Board of Review* [2000] 2 NZLR 9 (CA) an *R v Hansen* [2007] NZSC 7, [2007] 3 NZLR 1 at [92]-[98].
[391] *Wall v Fairfax New Zealand Ltd* [2018] NZHC 104 at [13] and Human Rights Law Centre "New Zealand High Court finds insulting cartoons did not breach hate speech legislation" (2018) < https://www.hrlc.org.au/human-rights-case-summaries/2018/6/1/new-zealand-high-court-finds-insulting-cartoons-did-not-breach-hate-speech-legislation>.
[392] Refer to Chapter 2, section 2.3.
[393] David Seymour and Andrew Little "Freedom of speech: Do we need to update our Human Rights Act?" (2019) <https://www.stuff.co.nz/national/politics/opinion/113785976/freedom-of-speech-do-we-need-to-update-our-human-rights-act>.

seeking repeal. As Geddis observes, New Zealand political culture strongly favours free expression, often treating restrictions as constitutionally suspicious.[394]

The Royal Commission of Inquiry found New Zealand's laws inadequate for addressing hate speech, recommending new offences that would criminalise "explicit and implicit calls for violence" or conduct that "normalises hatred." Importantly, it rejected adopting a UK-style "stirring up religious hatred" offence as unworkable but urged a high threshold that still captured coded and normalising speech.[395] This reflects an attempt to reconcile freedom of expression with the protection of dignity and social cohesion. The Royal Commission also stressed the relevance of Te Tiriti o Waitangi in shaping incitement law, noting that Māori communities, including takatāpui, remain disproportionately harmed by hate speech. This highlights the limits of New Zealand's current framework, which privileges expressive liberty over substantive equality.[396]

From a theoretical perspective, Mill's harm principle explains this reluctance: New Zealand law presumes expression should be curtailed only where direct, tangible harm can be shown. Yet this narrow standard ignores the collective harms of online hate speech, which are more accurately captured by alternative frameworks. Breyer's theory of active liberty suggests that democracy depends on inclusive participation, which can be undermined by targeted vilification. Waldron's focus on dignity and Fredman's emphasis on substantive equality reveal why existing proportionality analysis is insufficient: it overlooks the symbolic and cumulative harms that erode social cohesion and equal citizenship.

---

[394] Andrew Geddis "The State of Freedom of Expression in New Zealand: An Admittedly Eclectic Overview" (2008) 11 Otago L Rev 657 at 660-662

[395] Royal Commission of Inquiry Into The Terrorist Attack On Christchurch Mosques On 15 March 2019 "Part 9 - Social cohesion and embracing diversity" (2021) <https://christchurchattack.royalcommission.nz/the-report/part-9-social-cohesion-and-embracing-diversity/hate-crime-and-hate-speech> at 50-51

[396] Refer to Chapter 2, section 2.3

Balancing freedom of expression and hate speech regulation in New Zealand has therefore produced a skewed outcome. Courts and Parliament have prioritised liberty over equality, setting thresholds so high that hate speech laws are rarely enforced. Behavioural insights in Chapter 2 further demonstrate that online hate speech spreads rapidly through disinhibition and social learning, meaning that delays in recognition exacerbate harm. The challenge for reform is not whether freedom of expression should remain a cornerstone, but whether it should continue to trump the dignity and equality of vulnerable groups. As the Royal Commission concluded, hate speech undermines social cohesion and the assurance of belonging.[397] A more balanced framework would treat regulation as part of protecting democracy, not as a threat to it.

4.8 Conclusion

Hate speech incites harm onto the marginalised while the action in the real-world against an individual or group of people is a hate crime. Before the Christchurch shootings, Netsafe published data pertaining to harmful digital communications, and it was found that racial abuse online was widespread towards Asian, Māori and Pacific peoples.[398] However, there remains a notable absence of robust post-attack data. As the New Zealand Law Society noted in its 2024 submission on the Law Commission's issues paper, police only began flagging hate-motivated crimes in their system from 2019, and comprehensive data collection across the justice system remains insufficient.[399] This lack of systematic tracking obscures the true scope of hate-

[397] Royal Commission of Inquiry Into The Terrorist Attack On Christchurch Mosques On 15 March 2019 "Part 9 - Social cohesion and embracing diversity" (2021) <https://christchurchattack.royalcommission.nz/the-report/part-9-social-cohesion-and-embracing-diversity/hate-crime-and-hate-speech> at 50.
[398] Boundy, above n 7.
[399] Greta Yeoman "Too little known about hate-motivated offending - NZLS" (online ed, 2 April 2025) < https://www.capitalletter.co.nz/news/hate-crime/815799/too-little-known-about-hate-motivated-offending-nzls?utm_source=newsletter&utm_medium=email&utm_campaign=capital-letter-newsletter >

motivated offences and undermines efforts to hold perpetrators accountable. These institutional blind spots highlight both the persistent vulnerability of minority communities and the pressing need for reforms that strengthen data collection, enforcement consistency and legal protections.

This chapter has critically evaluated New Zealand's legislative landscape on hate speech. Across the New Zealand Bill of Rights Act 1990, the Human Rights Act 1993, the Harmful Digital Communications Act 2015, the Summary Offences Act 1981, the Broadcasting Act 1989, the Films, Videos, and Publications Classification Act 1993, and the Crimes Act 1961, the law provides partial tools but no coherent framework. Each statute reflects a different theoretical lens introduced in Chapter 2: NZBORA's proportionality jurisprudence resonates with Mill's liberty-based harm principle; the HRA nods toward Waldron's concern with dignity but applies it narrowly; the HDCA echoes Fredman's focus on substantive equality but remains confined to interpersonal harm; and the Classification Act illustrates Lessig's insight into "architecture," yet without adapting to the amplification effects of platform design. Read together, these statutes reveal how the liberty-first model dominates, while equality and dignity-based frameworks remain underdeveloped.

While the statutes acknowledge harm in certain contexts, they do so in a fragmented and largely reactive manner. The absence of explicit duties on digital platforms to prevent amplification of hate content leaves enforcement weak, creating what may be described as a *lex imperfecta*: a framework that exists in law but lacks substantive effect. In practice, New Zealand's regulatory system risks functioning as an extensive architecture with limited teeth; present in form, but not in deterrent force. As Chapter 3 also showed, New Zealand has leaned on soft law measures such as the Christchurch Call, which encourage voluntary cooperation from platforms. While symbolically significant, these initiatives underscore the limits of a framework that lacks

statutory duties of care. Without enforceable obligations, regulation risks remaining *lex imperfecta*..

The analysis in this chapter reinforces the need for a statutory duty of care and a consolidated legal framework, one that balances freedom of expression with proactive safeguards for dignity and equality. A more coherent approach is required to respond to systemic, coded, and digitally amplified hate speech that existing statutes fail to capture. The next chapter will therefore examine international approaches, considering how regulatory innovations overseas might inform a more effective, principled, and future-proof response for Aotearoa New Zealand.

Chapter 5: Comparative and Transnational Responses

The purpose of the comparative analysis is not to identify a single transferable regulatory model but to extract institutional design principles capable of adaptation within the New Zealand constitutional framework.

Part A: Criminalisation and Substantive Hate Speech Law

5.1 Introduction

This chapter is important to the thesis because it illustrates how supranational and national frameworks manage hate speech, offering comparative lessons for reform in New Zealand. By examining the European Union, as well as the national experiences of France and Germany, the chapter explores both substantive criminalisation and enforcement mechanisms, setting the stage for evaluating how New Zealand's framework can be improved.

This chapter examines how other jurisdictions and regional bodies have responded to online hate speech and harmful content. It does not aim to provide an exhaustive survey of international practice. Instead, it focuses on selected examples that illustrate two different levels of regulation. By comparing these frameworks, the chapter seeks to identify lessons applicable to New Zealand's context and evaluate their relevance to the central research question: How effective is New Zealand's legal framework in addressing hate speech on social media?

The first level is substantive criminalisation, where hate speech is prohibited directly through law. These measures rely on courts and penalties to deter harmful expression. The challenge here is to determine when expression should be classified as hate speech rather than as offensive but lawful speech. As argued in Chapter 2, this requires a careful balance between freedom of expression and the protection of dignity, equality, and democratic participation. Theories of free speech developed by Mill, Waldron, Meiklejohn, and Breyer, for example are particularly useful in explaining why limits on speech may sometimes be justified.

The second level is enforcement through regulatory and technical mechanisms. These measures assume that hate speech is already unlawful but ask how such laws can be applied in practice, especially in digital spaces. Here the focus shifts from individual speakers to online platforms and the tools they use to manage harmful content. Regulatory theories, such as Lessig's modalities and Murray's model of symbiotic regulation introduced in Chapter 2, highlight why criminalisation alone is insufficient. Effective responses must also address how platform design, economic incentives, and content moderation practices shape the spread of harmful expression.

The chapter is divided into two parts. Part A considers the substantive criminalisation of hate speech and its interaction with freedom of expression. It begins with international human rights law, before turning to the European Union, and finally to the national approaches of France and Germany. Part B examines enforcement and regulatory mechanisms. It looks at the Digital Services Act, methods of content moderation such as geo-blocking and takedown, and hybrid initiatives such as the Christchurch Call. The chapter concludes by drawing lessons for New Zealand, emphasising that criminal law and platform regulation must be seen as complementary rather than competing approaches.

5.2 Comparative Law as a Regulatory Methodology

Before examining these comparative frameworks, it is important to outline the methodological approach that guides the analysis in this chapter. This section introduces the comparative legal approach used in this thesis. While the core of this thesis examines New Zealand's legal responses to online hate speech and harmful content, it also draws on insights from international jurisdictions. These comparative references may appear diverse, they have been selected because they each represent distinct legal strategies for regulating digital harm. The aim is not to transplant foreign laws into the New Zealand system. Rather, it is to understand how other countries have structured legal and regulatory responses to online hate and to assess whether aspects of these models could inform local reform.

In exploring these foreign frameworks, this thesis follows what W. J. Kamba calls a contextual approach to comparative law. Kamba argues that legal rules must be interpreted in light of the historical, social, and institutional background of the legal system in which they exist. Without this context, comparisons may be misleading or superficial. He warns against the assumption

that laws can be copied or applied universally. As he notes, "it is wrong to regard the rules of law in different countries as automatically functionally equivalent.".[400] This caution is especially relevant for hate speech laws, which are shaped by national histories of free expression, minority protection, and political culture.

Similarly, Geoffrey Samuel highlights that comparative analysis must account for legal epistemology. This refers to the underlying concepts and ways of thinking that structure legal reasoning in different jurisdictions. For example, while New Zealand's legal system is rooted in common law reasoning and parliamentary supremacy, countries like Germany or France operate under civil law systems with different sources of authority and methods of interpretation. Samuel explains that without recognising these differences, comparative law risks becoming a "mere catalogue of foreign rules" instead of a meaningful analytical tool.[401] It asks the researcher to reflect on how legal concepts are formed and whether they can be transferred between jurisdictions with different institutional histories, political values, or cultural norms.[402] This attention to context is especially important in debates over hate speech, which touch on freedom of expression, group identity, and social cohesion; values that may carry different weight across legal cultures.

In addition, Jaakko Husa adds to this discussion by advocating for what he terms a legal-cultural methodology. Husa critiques purely functional or doctrinal approaches to comparison. He argues instead that effective comparative law must take account of the broader cultural and institutional logic that shapes legal systems.[403] He describes comparative law as a kind of

[400] W.J. Kamba, 'Comparative Law: A Theoretical Framework' (1974) 23(3) International and Comparative Law Quarterly 485, at 487.
[401] Geoffrey Samuel, An Introduction to Comparative Law: Theory and Method (Hart Publishing 2014) 31.
[402] Samuel, above n 401, at 32-34.
[403] Jaakko Husa, "About the Methodology of Comparative Law: Some Comments Concerning the Wonderland" (2015) 59 Revue Internationale de Droit Comparé 504 at 506.

"travelling in a wonderland".[404] Legal rules gain meaning only when placed within their native legal and cultural landscape. This approach supports the thesis's methodological stance. Although legal provisions from the UK or Germany may provide models for regulating online hate, their value for New Zealand depends on how well they align with local values, institutions, and constitutional norms.

Husa and Samuel both emphasise the importance of interdisciplinary thinking in comparative legal work.[405] Law does not operate in a vacuum. Comparisons are more insightful when they are connected with other perspectives, such as psychology, communication theory, or regulatory design. This justifies the thesis's broader framework, which combines legal comparison with theoretical models from behavioural science and regulatory theory.

This is also the reason why the thesis includes examples from several regions that may not seem like obvious comparators for New Zealand at first. These jurisdictions have been chosen because they are actively developing new legal responses to online hate speech and harmful content, and platform regulation. Even though their legal systems and political cultures differ from New Zealand's, their laws offer useful examples of how different governments are trying to deal with similar problems. By studying these international approaches, the thesis aims to understand what kinds of legal regulatory solutions might be adapted to New Zealand's own legal environment. This comparative perspective is not about replicating other countries' approaches but about learning from their experiences and assessing what could work in our context.

[404] Husa, above n 403, at 504.
[405] Husa, above n 403, at 516; Samuel, above n 398, at 41.

At a theoretical level, comparative law complements the regulatory perspective already discussed in this chapter. While regulatory theory helps to explain the interaction between law, platform design and social norms, comparative analysis shows how different legal systems are trying to regulate these interactions in practice. For example, the United Kingdom's Online Safety Act reflects a multi-modal regulatory approach that draws on legal duties, technical standards and user protection frameworks. By placing such examples alongside New Zealand's current approach, the thesis uses comparative law to assess not only what the law says, but how it works in context and whether alternative models may offer better outcomes.

Even though comparative law is useful for understanding how other countries deal with similar problems, it is not without difficulty. Many scholars have warned that comparing laws across countries can be misleading if we do not also consider the social, historical, and institutional differences that shape how those laws work. Husa, for example, points out that legal comparisons can become biased if they assume that Western legal ideas apply everywhere in the same way. Samuel also explains that even if two legal systems use similar words, they might still understand and apply those words very differently.

Taken together, these methodological insights form the foundation of the thesis's comparative analysis. The jurisdictions selected are not random. They have been chosen for the specific legal strategies they offer. Some, like the UK, provide recent statutory examples. Others, like Germany, offer constitutional traditions that balance speech rights with dignity and anti-discrimination protections. These comparisons are not meant to suggest direct transplantation. Rather, they enrich the evaluation of New Zealand's current and proposed legal framework. This comparative methodology underpins the discussion that follows, which examines international, European, and national approaches to the regulation of online hate speech.

5.3 Freedom of Expression and International Human Rights Law

New Zealand's obligations under international law arise both from treaties such as the ICCPR and from customary international law, principles reflected in instruments like the Vienna Convention on the Law of Treaties.[406] These sources guide domestic law and policy and provide a foundation for evaluating restrictions on expression.

The issue of cyber hate demonstrates how speech online can translate into harm in the real world. It intimidates vulnerable individuals and can silence their participation in public life, especially when it promotes hostility, discrimination, or violence.[407] This shows that freedom of expression is not an absolute right. The Human Rights Committee's General Comment No 34 on Article 19 of the ICCPR makes clear that expression must be balanced against other rights.[408] Although General Comments are not legally binding, they are widely accepted as authoritative interpretations of treaty provisions. They influence both international practice and domestic courts. General Comment No 34 allows states to impose restrictions on expression when these are lawful, necessary, and proportionate, particularly to prevent advocacy of hatred that amounts to incitement.[409] This reflects the tension highlighted in Chapter 2 between the liberty to speak and the protection of dignity, equality, and democratic participation.

[406] Statute of the International Court of Justice, above n 317; Law Commission *A New Zealand Guide to International Law and its Sources* (NZLC R34, 1996) at 16.
[407] United Nations General Assembly Report of the Special Rapporteur on the promotion and protection of the freedom of opinion and expression UN Doc A/74/486 (9 October 2019) at [56].
[408] ICCPR, above n 108; Human Rights Committee *General Comment No 34, Article 19: Freedoms of opinion and expression* CCPR/C/GC/34 (12 September 2011).
[409] ICCPR, above n 108, art 20; Human Rights Committee *General Comment No 31* (2004) at para 4.

The UN Special Rapporteur on freedom of expression has also emphasised that Article 19(3) sets strict limits on restrictions.[410] Laws must meet the principles of legality (clear and public rules), legitimacy (serving one of the listed aims in the Covenant), and necessity and proportionality (showing restriction is essential and narrowly applied). These tests are designed to prevent arbitrary state power. These standards are especially important in the online context. They provide a framework for deciding when hate speech crosses the line from offence to unlawful harm.

Scholars have underlined both the value and the limits of General Comment No 34. Kay stresses the importance of the tripartite test of legality, legitimacy, and necessity/proportionality, especially in the digital age where harmful content spreads quickly.[411] O'Flaherty notes the contested legal status of General Comments but argues they remain persuasive for courts interpreting domestic law.[412] Alfred de Zayas, reviewing Schabas' discussion on the *Faurisson v France* case, notes that "memory laws" criminalising certain historical claims may overstep the limits of proportionality and risk chilling legitimate academic or political speech.[413] Together these views show how the tripartite test functions in practice and how fragile the balance between harm prevention and freedom can be.

Proportionality analysis in comparative law reinforces these points. In *R v Oakes*, the Canadian Supreme Court developed a test requiring that (1) the objective of the restriction be sufficiently important; (2) the means be rationally connected to that objective; (3) the impairment of rights

[410] United Nations General Assembly Report of the Special Rapporteur on the promotion and protection of the freedom of opinion and expression UN Doc A/74/486 (9 October 2019) at [56].
[411] David Kay cited in Evelyn Mary Aswad (2020) 609 W&L L Rev 618.
[412] Michael O'Flaherty "Freedom of Expression: Article 19 of the International Covenant on Civil and Political Rights and the Human Rights Committee's General Comment No 34" (2012) HRLR 627 at 645
[413] Alfred de Zayas "W.A. Schabas, Nowak's CCPR Commentary: U.N. Covenant on Civil and Political Rights" Neth Int Law Rev 67 at 558.

be minimal; and (4) the benefits outweigh the harm caused by the restriction.[414] Similarly, Barak has argued that proportionality requires a proper purpose, rational connection, least restrictive means, and fair balance between interests.[415] Breyer's account of "active liberty" supports this framework by ensuring restrictions do not exclude participation in democratic life. These comparative perspectives confirm that proportionality is both a legal and normative principle for assessing restrictions on expression.[416]

These principles offer a structured way to judge laws that restrict expression. Applied to online hate speech, they ensure that regulation is used to protect dignity and equality without suppressing dissenting or minority voices. Mill's concern about overreach, Waldron's focus on the assurance of dignity, and Meiklejohn's emphasis on inclusive participation help explain why proportionality remains the best standard for evaluation.[417]

In conclusion, international human rights law establishes strict but flexible limits on expression. The tripartite test of legality, legitimacy, and necessity/proportionality provides a baseline against which national and regional laws on hate speech can be assessed. These benchmarks are essential for evaluating the European Union and national approaches in the next sections, and for how New Zealand's domestic laws align with international expectations.

5.4 The EU Framework Decision and the ECHR

[414] *R v Oakes* [1986] 1 SCR 103.
[415] Aharon Barak *PART IV: Proportionality Evaluated*. In Proportionality, Vol. Series Number 2. United Kingdom: Cambridge University Press, 2012 at 459.
[416] Breyer, above n 122, at 10.
[417] Waldron, above n 86 at 5 and Mill, above n 99 at 80.

The European Union has competence to harmonise criminal law where consistency is required across Member States, particularly in areas with cross-border impact.[418] This competence has been used to establish minimum standards for criminalising racist and xenophobic expression through the 2008 Council Framework Decision.[419] It reflects a shared commitment among Member States to prohibit hate speech while respecting constitutional traditions and freedom of expression, consistent with Article 2 of the Treaty on European Union.[420]

5.4.1 The 2008 Council Framework Decision: Combating Racism and Xenophobia

The Framework Decision establishes minimum standards for criminalising racist and xenophobic conduct across the EU.[421] While the scope of the Decision encompasses a range of acts, including incitement to violence and hate-motivated crimes, this section concentrates specifically on the provisions that pertain to expressive acts, such as hate speech.

Article 1 of the Framework Decision requires Member States to criminalise the "public incitement to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin."[422] It also addresses denial or gross trivialisation of genocide and crimes against humanity.. These expressive offences are directly relevant to this thesis, as they fall within the category of hate speech that many jurisdictions, including New Zealand, seek to regulate. Importantly, the Framework Decision does not aim for full harmonisation. Instead, it allows for flexibility in

[418] European Commission, above n 355.
[419] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, art 6.
[420] Consolidated version of the Treaty on European Union (26 October 2010) *Official Journal of the European Union C 326/15,* art 2.
[421] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, art 6.
[422] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, art 1.

implementation while affirming a shared commitment to restrict hate-based incitement at the supranational level.[423]

The preamble provides guidance relevant to online expression. Preamble 6 recognises the growing prevalence of harmful acts within information systems, while Preamble 8 interprets "religion" broadly to include convictions and beliefs.[424] This extends protection beyond race and ethnicity to cover religiously targeted abuse, which is significant given the rise of online Islamophobia and antisemitism. From a theoretical perspective, the prohibition of incitement to hatred reflects Waldron's concern with protecting dignity and social assurance, while provisions on genocide denial raise Mill's warning against overly broad restrictions.[425] As Henrard notes, the EU's approach to protecting religious minorities has been cautious and uneven, which makes the broad reading in Preamble 8 especially noteworthy.[426]

From New Zealand's perspective, the Framework Decision provides a useful benchmark. It shows how a supranational instrument can require action against incitement while preserving national variation, and it illustrates the importance of expanding protected grounds such as religion, which remain underdeveloped in New Zealand's current regime.

5.4.2 Penalty Management and Enforcement

[423] Laurent Pech, "The Rule of Law as a Constitutional Principle of the European Union" (2010) 6 EuConst 359 at 370.
[424] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, preamble 6 and 8.
[425] Waldron, above n 3 at 5 and Mill, above n 99 at 80.
[426] Kristin Henrard, "EU Law's Half-Hearted Protection of Religious Minorities" (2021) 12 Religions 830 at 5-6.

The Council Framework Decision also outlines penalties for individuals and legal entities involved in acts of racism and xenophobia. These penalties extend beyond criminal sanctions to include non-criminal measures such as the denial of public benefits or aid, temporary or permanent exclusion from commercial activities, judicial supervision, or even judicial winding-up orders. These provisions demonstrate the EU's commitment to creating a multifaceted enforcement regime that holds both individuals and organisations accountable. While Article 1 offences focus on incitement and denial of genocide, the penalty framework demonstrates the EU's commitment to accountability at both the individual and organisational level.

Implementation, however, has been uneven. As Gagliardone et al. observe, some states integrated the Framework Decision fully, while others adopted only partial measures, particularly concerning online enforcement.[427] This unevenness illustrates Breyer's proportionality principle: even when hate speech prohibitions exist, enforcement must remain necessary and balanced.

The European Convention on Human Rights adds a further layer. Article 10 protects freedom of expression but permits restrictions "necessary in a democratic society." Strasbourg jurisprudence, including *Handyside v United Kingdom* and *Garaudy v France*, confirms that hate speech restrictions must still pass a proportionality test under Article 10(2).[428] This means Member States cannot rely solely on the Framework Decision but must implement it in a way

[427] Iginio Gagliardone and others Countering online hate speech (UNESCO, 2015) at 37- 40.
[428] *Handyside v United Kingdom* (1976) 1 EHRR 737 (ECHR) at [49]–[50]; *Garaudy v France* (2003) V ECHR 383 (ECHR, App No 65831/01) at [27].

consistent with the ECHR. As Szyszczak observes, this creates a complex relationship requiring Member States to reconcile EU criminal standards with ECHR oversight.[429]

From a New Zealand perspective, the Framework Decision offers two lessons. It demonstrates how a regional body can require action against incitement while preserving national variation. It also highlights the role of proportionality review to ensure that restrictions on expression protect dignity without unduly suppressing debate. These lessons will be revisited in section 5.8.

5.5 National Approaches: France and Germany

This section now turns to two Member State case studies, France and Germany. Both have developed national laws directed at online hate speech, although their outcomes differ. Examining these examples provides deeper insight into how states balance enforcement with freedom of expression. It also shows how national approaches interact with supranational frameworks like the DSA, and what lessons may be relevant for New Zealand.

France and Germany have both attempted ambitious regulation of online hate speech. Yet their experiences diverge. France's 2019 Avia Law was struck down by the Constitutional Council as unconstitutional, while Germany's 2017 NetzDG remains in force, though subject to ongoing criticism. These contrasting examples illustrate the challenges of balancing effective legal tools against the dangers of overreach.

[429] Erika Szyszczak "Antidiscrimination Law in the European Community" (2009) 32 Fordham International Law Journal 623 at 631-633.

France introduced the Avia Law in 2019 to combat hate speech online, granting social media companies the authority to monitor and remove discriminatory content within twenty-four hours.[430] For terrorist or child exploitation material, the deadline was reduced to one hour.[431] Failure to comply exposed platforms to fines of up to €1.25 million. While the law aimed to address urgent harms, its strict timelines and broad mandates raised concerns about over-censorship.[432] Ultimately, the French Constitutional Court invalidated significant portions of the law, citing violations of freedom of expression and risks of overreach.[433]

The Constitutional Council reasoned that strict deadlines placed disproportionate pressure on platforms. In practice, companies would remove lawful content simply to avoid sanctions. This created a systemic risk of "overblocking" expression that was not unlawful, which the Council found incompatible with constitutional protections. This reflects Mill's concern that state-imposed restrictions can silence legitimate debate by encouraging private actors to censor more than is required.[434]

Although the Avia Law was ultimately struck down, its emphasis on rapid takedown of hate speech content highlights the difficulty of designing proportionate enforcement mechanisms. For New Zealand, the case demonstrates the risks of adopting overly strict timelines that incentivise over-removal, even where the objective of combating hate speech is legitimate.

[430] Chloe Hadavas "France's New Online Hate Speech Law Is Fundamentally Flawed" (2020) <https://slate.com/technology/2020/05/france-hate-speech-law-lutte-contre-haine-sur-internet.html>.
[431] Hadavas, above n 430.
[432] Hadavas, above n 430.
[433] Laura Kayali "French constitutional court strikes down most of hate speech law" (2020) <https://www.politico.eu/article/french-constitutional-court-strikes-down-most-of-hate-speech-law/>
[434] Laura Kayali "French constitutional court strikes down most of hate speech law" (2020) <https://www.politico.eu/article/french-constitutional-court-strikes-down-most-of-hate-speech-law/>

Germany has been at the forefront of regulating hate speech online, introducing the Network Enforcement Act (officially referred to as the Netzwerkdurchsetzungsgesetz or "NetzDG") in 2017. Enacted by the Bundestag, it was one of the first attempts worldwide to impose detailed obligations on major platforms. While similar to the Avia Law in its goals, its more structured design has allowed it to remain in force.[435] The law created a structured framework for platform accountability.

Article 1 of the NetzDG outlines a framework designed to enhance transparency in the handling of user complaints. To achieve this, the law imposes the following key obligations on social media platforms:

- *Reporting System*: Platforms must establish a clear, accessible, and user-friendly mechanism for reporting prohibited content.[436]
- *Content Evaluation*: Platforms are required to evaluate reported content promptly to determine whether it violates the law.[437]
- *Content Removal*: Upon receiving a valid complaint, platforms must remove or block access to the content within 24 hours for clearly unlawful content. For less obvious cases, platforms are given seven days to act.[438]
- *Regulated Self-Governance*: When complaints are referred to a recognised institution of regulated self-governance, the institution also has up to seven days to decide on the content.[439]

By mandating strict timelines and reporting requirements, the NetzDG aims to foster accountability among social media platforms while ensuring that harmful content does not

[435] Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act, NetzDG) - Basic Information (2017) 2021 Bundesministerium der Justiz und für Verbraucherschutz art. 1, section 1.
[436] NetzDG, above n 435 at 482.
[437] NetzDG, above n 435 at 482.
[438] NetzDG, above n 435 at 482.
[439] NetzDG, above n 435 at 482.

remain online indefinitely. Penalties for non-compliance can reach up to €50 million.[440] The Act applies to platforms with more than two million users in Germany. It has been both praised for improving accountability and criticised for potential overreach and censorship.[441]

The scope of the NetzDG is outlined in Section 1, extending its applicability to the German Criminal Code (Strafgesetzbuch - StGB).[442] Section 1(3) of the NetzDG incorporates provisions of the StGB to classify unlawful content into categories such as "endangering the democratic rule of law; treason and endangering external security; resistance to state authority; offences against public order; offences relating to religion and ideology; offences against sexual self-determination; insult; offences against personal liberty; and forgery of documents".[443] Of particular relevance to hate speech are the offences against public order, specifically addressed in Section 130 which criminalises incitement to hatred and violations of human dignity. Penalties range from three months to five years' imprisonment.[444]

It could be argued that the NetzDG risks exceeding the proportional limits of Article 19 of the ICCPR, which protects freedom of expression but permits necessary and proportionate restrictions.[445] Human Rights Watch, Zipursky, and Zurth have each argued that the Act incentivises platforms to over-remove content, effectively outsourcing censorship to private companies. [446] Tworek and Leerssen also caution that the law risks chilling lawful speech by

[440] NetzDG, above n 435 at 482.
[441] Patrick Zurth "The German NetzDG as Role Model or Cautionary Tale? Implications for the Debate on Social Media Liability" (2021) 31 Fordham Intellectual Property, Media and Entertainment Law Journal at
[442] German Criminal Code 1998.
[443] German Criminal Code 1998, ss 86, 86a, 89a, 91, 100a, 111, 126, 129 to 129b, 130, 131, 140, 166, 184b in connection with 184d, 185 to 187, 241 or 269 & Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act, NetzDG) - Basic Information (2017) art 1, s 1.
[444] German Criminal Code 1998, s 130 (1).
[445] International Covenant on Civil and Political Rights (opened for signature 16 December 1066, entered into force 23 March 1976), art. 19.
[446] Human Rights Watch "Germany: Flawed Social Media Law. NetzDG is Wrong Response to Online Abuse" 2018 < https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law>; Rebecca Zipursky "Nuts

imposing disproportionate burdens.[447] This tension reflects Breyer's emphasis on proportionality, where restrictions must be justified by a proper purpose and avoid imposing excessive burdens on expression.[448]

While Article 19 of the ICCPR affirms freedom of expression, it also allows restrictions that are lawful, necessary, and proportionate. Critics contend that the NetzDG, though intended to combat hate speech, imposes obligations that risk exceeding these limits by encouraging platforms to remove content too broadly.[449] This tension highlights the broader challenge of reconciling robust content moderation with the protection of fundamental freedoms of expression in a democratic society.[450] This debate reflects Breyer's emphasis on proportionality: even where regulation has a proper purpose, it must not impose excessive burdens on expression.

A separate controversy highlights Germany's struggle to balance expression and dignity. In 2016, satirist Jan Böhmermann read a poem mocking Turkish President Recep Tayyip Erdoğan. The poem, which included inflammatory content mocking Erdoğan's personal attributes, sparked diplomatic outrage.[451] As a consequence of this, Erdoğan pressed the German government to initiate criminal proceedings under Section 103 of the German Criminal

About NETZ: The Network Enforcement Act and Freedom of Expression- " (2019) 42 Fordham International Law Journal 1325 at 1331.; and Patrick Zurth "The German NetzDG as Role Model or Cautionary Tale? Implications for the Debate on Social Media Liability" (2021) 31 Fordham Intellectual Property, Media and Entertainment Law Journal.

[447] Heidi Tworek and Paddy Leerssen *An Analysis of Germany's NetzDG Law* (Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, 15 May 2019).

[448] Breyer, above n 122, at 10.

[449] Human Rights Watch "Germany: Flawed Social Media Law. NetzDG is Wrong Response to Online Abuse" 2018 < https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law>

[450] Rebecca Zipursky "Nuts About NETZ: The Network Enforcement Act and Freedom of Expression- " (2019) 42 Fordham International Law Journal 1325 at 1331.

[451] Deutsche Welle (www.dw.com) "German satirist Jan Böhmermann sues Angela Merkel over Erdogan poem remark | DW | 02.04.2019" (2021) <https://www.dw.com/en/german-satirist-jan-böhmermann-sues-angela-merkel-over-erdogan-poem-remark/a-48158329>.

Code, a provision prohibiting insults against representatives of foreign states. This action, sanctioned by then-Chancellor Angela Merkel, drew significant controversy for its perceived infringement on free expression.[452] Human Rights Watch and other advocacy groups criticized the move, urging Germany to repeal Section 103, which they argued was incompatible with Article 19 of the ICCPR. In response, Germany's Parliament eventually repealed the provision, demonstrating a commitment to aligning domestic law with international human rights standards.[453]

The Böhmermann case underscores how provisions designed to protect dignity and international relations can, in practice, be used to suppress satire and political critique. Its repeal shows Germany's willingness to recalibrate its laws when restrictions are no longer proportionate. In theoretical terms, the episode reflects Mill's concern that state power may be used to stifle robust debate, while also underscoring Waldron's point that dignity and assurance remain central in setting the limits of acceptable expression.[454]

The legal debate often turns on whether the restrictions in Section 130 of the German Criminal Code, which criminalises incitement to hatred and violence, remain proportionate under Article 19 of the ICCPR. Critics have argued that the NetzDG amplifies these restrictions in ways that risk overreach. The repeal of Section 103 StGB after the Böhmermann case (discussed above) shows that Germany is willing to recalibrate when restrictions on expression are no longer defensible under international standards.

[452] Zipursky, above n 450, at 1331.
[453] Philip Oltermann "Obscure German law gives Angela Merkel a diplomatic headache" (2016) <http://www.theguardian.com/world/2016/apr/14/obscure-german-law-angela-merkel-recep-tayyip-erdogan>. And Zipursky, above n 450 at 1328.
[454] Mill, above n 43 at 80; and Waldron, above n 3, at 34

Taken together, these examples demonstrate a distinctive balance: Germany retains strong prohibitions on incitement to hatred under Section 130 but also shows sensitivity to the risks of excessive censorship by discarding outdated provisions. In theoretical terms, this reflects Breyer's proportionality framework, which requires that restrictions pursue a proper purpose, remain rationally connected to that purpose, and avoid imposing excessive burdens on expression.[455] The German approach illustrates how proportionality analysis is not just abstract doctrine but a practical tool that shapes the boundaries of hate speech regulation. For New Zealand, it highlights how courts and legislators might assess whether restrictions in the Human Rights Act or proposed reforms genuinely strike the balance between protecting dignity and preserving open democratic debate.

The principle that human dignity is inviolable is enshrined in Chapter 1 of the Fundamental Rights of the European Union and serves as a cornerstone of Germany's constitutional framework. [456] Germany's Federal Constitutional Court has repeatedly affirmed that human dignity is the "untouchable" basis of the Basic Law, shaping how all other rights, including freedom of expression, are interpreted. While Germany criminalises hate speech, this does not eliminate free expression. Instead, courts consistently frame expression as subject to limits where it undermines dignity or constitutes incitement to hatred. This approach reflects Waldron's theory that the assurance of dignity for vulnerable groups is itself a democratic good. For Waldron, the visibility of hate speech signals to targeted groups that they are not full members of society; Germany's dignity-based framework provides one response to that risk.[457]

[455] Breyer, above n 122, at 33.
[456] Charter of Fundamental Rights of The European Union 2012 European Union [2012] OJ 2012/C 326/02
[457] Waldron, above n 3, at 94.

For New Zealand, where the statutory framework under the Human Rights Act has been criticised as under-inclusive, Germany's emphasis on dignity highlights a different constitutional starting point: protection of social assurance can sometimes justify narrower boundaries for speech.

The NetzDG requires social networks to ensure transparency in decision-making by informing users of the reasons behind content removal or other moderation actions. This process is integral to effectively controlling illegal content while maintaining a semblance of accountability.[458] Importantly, the Act applies not only to social media platforms headquartered in Germany but also to those operating outside its borders, including major players based in the United States and China. As highlighted by the German Ministry of Justice and Consumer Protection, this extraterritorial application underscores Germany's commitment to regulating online hate speech and illegal content on platforms accessible within its jurisdiction.[459] However, it also raises questions about enforceability and the potential for conflicts with other nations' legal frameworks, particularly those with less restrictive approaches to free speech regulation.

Despite its aim of improving accountability, the NetzDG has sparked critical debates. Messaging services are excluded, enforcement powers are broad, and penalties are severe. Tworek and Zipursky argue that while it enhances accountability, it risks encouraging excessive caution by platforms.[460] On the other hand, Keller, warn that systemic duties of care

[458] Bundesministerium der Justiz und für Verbraucherschutz "FAQ: Act to Improve Enforcement of the Law in Social Networks, 2017" (2021) <https://www.BMJV.de/SharedDocs/FAQ/EN/NetzDG/NetzDG.html>.
[459] Federal Ministry of Family Affairs and Federal Ministry of Justice and Consumer Protection (2017) <https://www.bmjv.de/SharedDocs/Pressemitteilungen/DE/2017/03142017_Monitoring_SozialeNetzwerke.html>
[460] Heidi Tworek and Paddy Leerssen *An Analysis of Germany's NetzDG Law* (Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, 15 May 2019).

can drive over-censorship.[461] These critiques again raise the question of proportionality, and whether Germany's model strikes the fair balance between dignity and free expression.

The Act applies only to large platforms with more than two million users, leaving out messaging services such as WhatsApp and professional networks like LinkedIn and Xing. Scholars have noted that these private but large-scale spaces are often hubs for harmful content, creating a regulatory gap.[462]

In addition, the Act requires major platforms to submit annual transparency reports.[463] These reports must detail the number and nature of complaints received, the rationale for their decisions, and the composition of their decision-making teams. These records must also be publicly accessible, a measure designed to enhance corporate accountability and maintain public trust.

From a theoretical standpoint, the NetzDG illustrates how Lessig's "law" and "architecture" modalities combine: legal mandates shape the technical systems of reporting and takedown, but critics warn that such structural incentives may over-regulate speech.[464] Proportionality review therefore remains central in evaluating whether these mechanisms protect dignity without creating excessive burdens on expression.

Part B: Enforcement, Platforms and Technical Regulation

[461] Daphne Keller "Broad Consequences of a Systemic Duty of Care for Platforms" (1 June 2020) The Center for Internet and Society <https://cyberlaw.stanford.edu/blog/2020/06/broad-consequences-systemic-duty-care-platforms.>
[462] Ben Knight "Germany implements new internet hate speech crackdown" DW.COM (1 January 2018) < https://www.dw.com/en/germany-implements-new-internet-hate-speech-crackdown/a-41991590>.
[463] Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act, NetzDG) - Basic Information (2017)
[464] Lessig, above n 84 at 664; and Cohen, above n 97, at 5.

5.6 The EU Digital Services Act (DSA)

The Digital Services Act (hereinafter "DSA"), enacted on November 16, 2022, marks a landmark development in the European Union's efforts to regulate online platforms and protect digital rights. Its core objectives are to "keep users safe from illegal goods, content, or services, and to protect their fundamental rights online."[465] In practical terms, this means strengthening transparency and accountability in how platforms moderate content, while giving users more control. From a theoretical perspective, the DSA reflects Lessig's idea that law can shape the "architecture" of the internet, and Breyer's concern with proportionality in ensuring that regulatory burdens are balanced against free expression.[466]

5.6.1 Key Legislative Context

The DSA was introduced under the authority of the European Commission, as provided by Directive (EU) 2015/1535, which set out procedures for technical regulations and Information Society services.[467] President Ursula von der Leyen identified the DSA as a political priority, presenting it as central to creating a coherent legal framework for online platforms across the EU.[468] The DSA reflects a growing recognition of the need for robust regulations to address evolving challenges in cyberspace, particularly the proliferation of harmful and illegal content.

---

[465] Council of the EU "What is illegal offline should be illegal online: Council agrees position on the Digital Services Act" (25 November 2021) <https://www.consilium.europa.eu/en/press/press-releases/2021/11/25/what-is-illegal-offline-should-be-illegal-online-council-agrees-on-position-on-the-digital-services-act/>

[466] Lessig, above n 84 at 664; & Breyer, above n 122, at 22.

[467] Christoph Schmon and Paige Collings "The Adoption of the EU's Digital Services Act: A Landmark Year for Platform Regulation: 2022 in Review" (2022) Electronic Frontier Foundation <https://www.eff.org/deeplinks/2022/12/adoption-eus-digital-services-act-landmark-year-platform-regulation-2022-year> and Directive (EU) 2015/1535 Of The European Parliament And Of The Council 2015 [2015] OJ L 241/1.

[468] European Commission "Digital Services Act – deepening the internal market and clarifying responsibilities for digital services" (2021) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12417-Digital-Services-Act-deepening-the-internal-market-and-clarifying-responsibilities-for-digital-services>.

5.6.2 Public Consultation and Foundational Principles

Between June and September 2020, the European Commission invited public consultation to gather evidence and opinions on the proposed legislative reforms. The consultation process reflected the EU's emphasis on democratic participation and inclusivity. The DSA builds upon the foundational pillars of the e-Commerce Directive, with key updates to strengthen the regulatory framework:

Country-of-Origin Principle: Online platforms are regulated primarily by the laws of their country of establishment, ensuring consistency while allowing exceptions for issues related to public policy, health, security, and consumer protection.[469]

Liability Exemptions: Hosting platforms are exempt from liability for illegal content if they lack "actual knowledge" of the content's illegality and act promptly upon gaining such awareness. [470] Article 14 paragraph 2 establishes clear criteria for this exemption, emphasising the importance of precise notice-and-action mechanisms.[471]

---

[469] Sally Broughton Micova and Alexandre de Streel *Digital Services Act – deepening the internal market and clarifying responsibilities for digital services* (Centre on Regulation in Europe, 2020).

[470] Micova and de Streel, above n 469, at 8.

[471] Proposal for a Regulation of The European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC.2021 , *Notice and actions mechanisms,* at Article 14. The elements in which providers must Article 14 paragraph 2 specifies what must be done to be exempted: (Unclear sentence)

"…

(a) an explanation of the reasons why the individual or entity considers the information in question to be illegal content;

(b) a clear indication of the electronic location of that information, in particular the exact URL or URLs, and, where necessary, additional information enabling the identification of the illegal content;

(c) the name and an electronic mail address of the individual or entity submitting the notice, except in the case of information considered to involve one of the offences referred to in Articles 3 to 7 of Directive 2011/93/EU;

(d) a statement confirming the good faith belief of the individual or entity submitting the notice that the information and allegations contained therein are accurate and complete"

Prohibition of General Monitoring Obligations: Member States cannot impose a general obligation on hosting platforms to monitor content proactively.[472]

Co- and Self-Regulation: The DSA encourages the development of codes of conduct to facilitate the implementation of its principles, promoting collaboration between platforms and regulatory bodies. [473] In other words, the notion of co- and self-regulation would help confront potential online harm; online hate speech being one of many.

5.6.3 Harmonising Online Content Moderation

The DSA represents a horizontal regulation, providing overarching guidelines applicable to all platforms and types of illegal content. It complements vertical regulations targeting specific areas, such as child sexual abuse materials, terrorism, and copyright infringement. By modernising these frameworks, the DSA seeks to address diverse forms of harmful content while ensuring proportionality and fairness in enforcement.

[472] Micova and de Streel, above n 469, at 9.
[473] Micova and de Streel, above n 469, at 9.

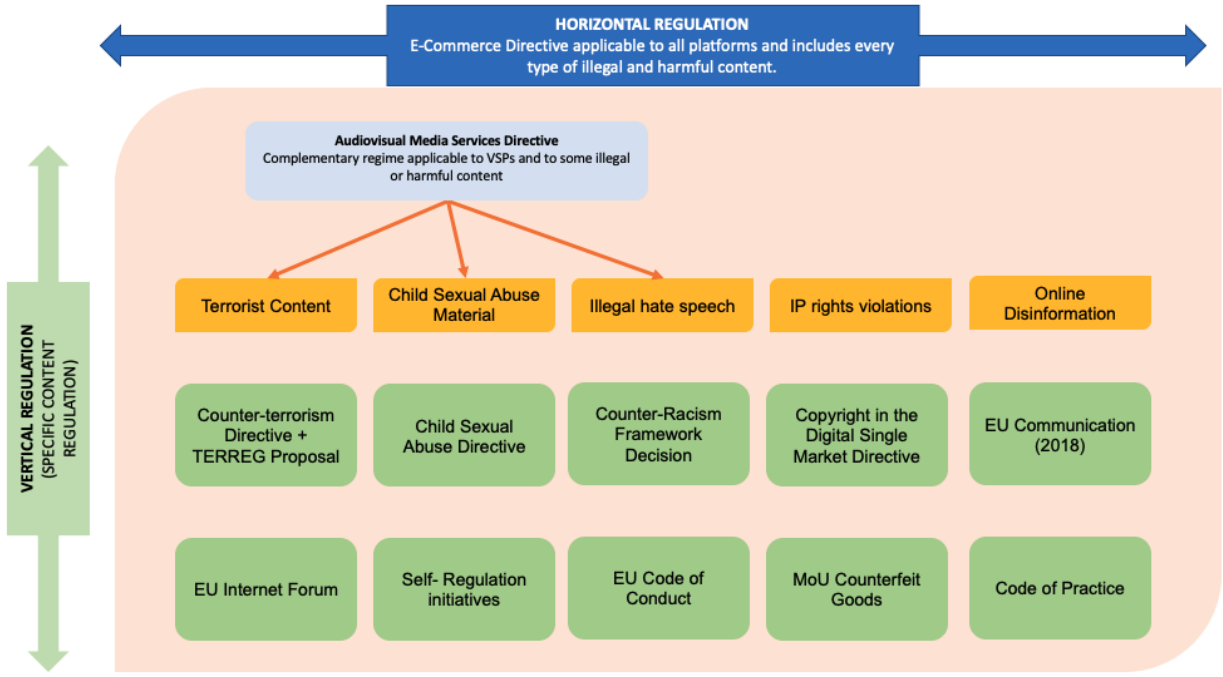

Figure 1: The EU regulatory framework for online content moderation. Adapted from Online Platforms' Moderation of Illegal Content Online - Law, Practices and Options for Reform. De Streel, et al.[474]

Figure 1 illustrates the EU regulatory framework for online content moderation, highlighting the interplay between horizontal and vertical regulations. Adapted from the report by de Streel et al., this framework underscores the importance of clear responsibilities for digital services in combating illegal content.

An overview of how the EU regulates illegal content highlights the overarching e-Commerce Directive, which serves as a horizontal regulation applying to all platforms and all types of illegal or harmful content. [475] In the context of hate speech, the Digital Services Act (DSA) address specific obligations for content moderation and platform accountability, demonstrating the EU's evolving commitment to clear and enforceable standards for digital services.

[474] Alexandre de Streel and others *Online Platforms' Moderation of Illegal Content Online - Law, Practices and Options for Reform* (European Parliament's committee on Internal Market and Consumer Protection, Scientific Foresight Unit (STOA), PE 656.318, February 2021) at 19.
[475] Micova and de Streel, above n 469, at 19.

An overview of the EU's approach to illegal content shows that the e-Commerce Directive functions as a horizontal regulation, applying to all platforms and all types of illegal or harmful material. In the area of hate speech, this framework is supplemented by the Digital Services Act (DSA), which introduces specific obligations for content moderation and platform accountability and other vertical regulation. These developments reflect the EU's growing commitment to clear and enforceable standards for digital services.

At this point, it is essential to critically assess the European Union's legislative framework governing illegal content, with particular emphasis on online hate speech. The EU's regulatory landscape encompasses several key directives and regulations aimed at addressing a spectrum of harmful materials. These include targeted measures such as directives on combating child sexual abuse materials and terrorism, alongside broader frameworks like the General Data Protection Regulation (GDPR) and the Audiovisual Media Services Directive (AVMSD). Additionally, the Digital Single Market (DSM) Copyright Directive (DSMD) addresses challenges related to copyright infringement and content moderation in the digital sphere. Together, these regulations underscore the EU's commitment to maintaining a safe and rights-based digital environment. Understanding the scope and interplay of these legislative instruments is critical to evaluating their effectiveness in addressing online hate speech while ensuring the protection of fundamental rights across the EU's jurisdiction.[476]

By virtue of the European Board for Digital Services (which is an independent advisory group for national authorities), enforcing the DSA will be under the purview of Member States as

[476] Micova and de Streel, above n 469, at 36.

opposed to the European Commission.[477] EU Member States are required to transpose the Council Framework Decision 2008/913/JHA into national legislation. Many Member States have implemented laws to align with the directive, although the degree and manner of implementation vary. For instance, Germany's legislation on incitement to hatred (Volksverhetzung[478]) criminalises hate speech targeting groups based on race, religion, or ethnicity, closely reflecting the directive's goals.[479] However, not all Member States have fully harmonised their laws with the directive, and enforcement mechanisms can differ significantly.[480] These variations illustrate the challenges of achieving uniformity across the EU while respecting the subsidiarity principle. From a theoretical perspective, this unevenness echoes Breyer's concern with proportionality: restrictions must be carefully tailored and justified across different legal systems.[481] It also recalls Meiklejohn's emphasis on democratic participation, as fragmentation may weaken the assurance that all citizens across the EU enjoy equal protection against hate speech.[482]

The DSA, alongside the Digital Markets Act (DMA), plays a pivotal role in reshaping the regulatory landscape of the digital economy within the European Union. While the DSA is concerned primarily with safeguarding user safety and ensuring transparency in online platforms, the DMA targets competition law, seeking to curb the monopolistic power of large

[477] Oliver Bell and Vito Petretti "European Commission Publishes the Digital Services Act" (29 December 2020) Lexology Tech & Sourcing @ Morgan Lewis <https://www.lexology.com/library/detail.aspx?g=913db9d9-b813-4520-b827-97365669b8f0>
[478] Criminal Code (Germany) (Strafgesetzbuch, StGB) translation at Federal Ministry of Justice and Consumer Protection <https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html.>
[479] European Commission Report from the Commission to the European Parliament and the Council on the implementation of Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law COM(2014) 27 final (27 January 2014).
[480] European Parliament Answer given by Ms Jourová on behalf of the Commission [E-001040/2018] (17 April 2018) <https://www.europarl.europa.eu/doceo/document/E-8-2018-001040-ASW_EN.pdf.>
[481] Breyer, above n 122, at 8 and 33.
[482] Meiklejohn, above n 104, at 26.

technology companies.[483] Together, these legislative instruments aim to create a cohesive digital environment that balances user protections with market fairness. From a theoretical standpoint, this dual framework reflects Lessig's and Murray's idea of regulation through multiple modalities: the DSA operates through "law" and "architecture" to shape platform practices, while the DMA reinforces "market" constraints by limiting anti-competitive conduct.[484] When considered together, they illustrate how layered regulation can be used to protect both individual rights and systemic fairness in the online sphere.

Notably, the majority of leading technology firms are headquartered in the United States, and their global influence has often outpaced regulatory frameworks.[485] The DSA and DMA respond to this imbalance by establishing a unified rulebook to harmonise platform operations across the EU. This harmonisation recalls earlier historical transitions, such as the introduction of traffic laws after the widespread adoption of automobiles, where standardisation was essential to ensure order and safety. In regulatory theory terms, this demonstrates how law can evolve to match technological disruption, aligning with Breyer's proportionality framework by seeking to strike a balance between innovation, market efficiency, and the protection of fundamental rights.

The introduction of the DSA during the global COVID-19 pandemic underscores the growing centrality of digital technologies to both social interaction and economic life.[486] Recognising this interdependence, the DSA proposal emphasises the need for a robust legislative framework

[483] Natasha Lomas "Understanding Europe's big push to rewrite the digital rulebook" (31 December 2020) <https://techcrunch.com/2020/12/30/understanding-europes-big-push-to-rewrite-the-digital-rulebook/>.
[484] Lawrence Lessig "The New Chicago School" (1998) 27 The Journal of Legal Studies 661 at 664.; and Murray, above n 134, at 240.
[485] Lomas, above n 483.
[486] Proposal for a Regulation of The European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC.2021, above n 376.

that delineates the responsibilities of social media platforms, intermediary services, and online marketplaces.[487] It raises the benchmark for accountability by mandating transparency in content moderation practices, advertising policies, and algorithmic processes, with specific duties to identify and mitigate systemic risks, including those linked to manipulative techniques.[488] From a theoretical perspective, this emphasis on systemic risk aligns with Lessig's account of "architecture" as a regulatory modality: by shaping the design of online platforms, the DSA seeks to embed safeguards that protect user rights while addressing structural drivers of harm.[489]

Central to the DSA is the principle of self-regulation, underpinned by mechanisms for the submission and handling of notices. Article 6 exempts hosting providers from liability for illegal content stored at a user's request where they lack "actual knowledge" of its illegality, or act promptly to remove or disable access once aware. This approach reflects an attempt to balance the operational realities of digital platforms with the need to regulate harmful content, including hate speech.[490] By emphasising "actual knowledge or awareness," the DSA narrows the circumstances under which liability arises, offering clarity for both operators and users. In theoretical terms, this reflects Breyer's concern with proportionality: liability is not imposed pre-emptively but only when platforms fail to act responsibly, thereby seeking to protect democratic participation without encouraging excessive censorship.

[487] Proposal for a Regulation of The European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. 335.
[488] Proposal for a Regulation of The European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. 335.
[489] Lawrence Lessig "The New Chicago School" (1998) 27 The Journal of Legal Studies 661 at 664.
[490] Regulation 2022/2065 of the European Parliament and of the on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJEU L 277/1.

The Council of Europe has also undertaken initiatives to combat hate speech through the establishment of the Committee of Experts on Combating Hate Speech (ADI/MSI-DIS). This committee is tasked with preparing comprehensive recommendations that address hate speech in online environments within the framework of human rights. Drawing from the case law of the European Court of Human Rights, Council of Europe texts, and the influence of campaigns like the No Hate Speech Movement Youth Campaign, the Committee provides a multi-faceted approach to tackling hate speech.[491] Its two subcommittees, the Steering Committee on Anti-Discrimination, Diversity, and Inclusion (CDADI) and the Steering Committee on Media and Information Society (CDMSI), focus on their respective areas of expertise to ensure that recommendations are both thorough and actionable.[492]

Currently, the Committee's draft recommendations are under review, with plans for deliberation and adoption by Member States.[493] The process reflects a collaborative approach that situates hate speech regulation within a framework of diversity, inclusion, and respect for human rights. In normative terms, this resonates with Waldron's theory of dignity: regulation is justified not simply to suppress harmful expression, but to assure vulnerable groups of their equal standing in society.[494] Unlike the Digital Services Act, which imposes binding obligations and penalties on platforms, the Council of Europe's framework emphasises dialogue, inclusion, and soft-law guidance.

This contrast highlights two complementary strands of regulation: the EU's reliance on enforceable duties to secure compliance, and the Council's use of normative principles to shape

[491] Committee of Experts on Combating Hate Speech (ADI/MSI-DIS) "Background document" Combating Hate Speech (ADI/MSI-DIS) (25 May 2020) Council of Europe <http://rm.coe.int/09000016809e90ea>.
[492] Committee of Experts on Combating Hate Speech (ADI/MSI-DIS), above n 491.
[493] Committee of Experts on Combating Hate Speech (ADI/MSI-DIS), above n 491.
[494] Waldron, above n 3, at 34 and 108.

standards through persuasion and consensus. Taken together, they reflect different ways of operationalising the balance between expression and dignity in the digital sphere. For New Zealand, the European experience suggests that effective regulation may require a blend of these approaches: binding duties to ensure accountability, and normative guidance to protect dignity and social assurance without overreliance on coercion.

5.7 Content Moderation Tools

This section examines platform-side tools that operationalise content rules: notice-and-action workflows, geo-blocking, filtering, takedown and account sanctions. Each tool manages risk differently. The core questions are accuracy, speed, due process, and effects on lawful speech. The DSA gives the clearest public-law frame for these choices. Lessig's "architecture" helps explain why design decisions matter, while Breyer's proportionality test provides the standard for judging when restrictions go too far.

Social media's rapid growth, with approximately 2.95 billion users globally, has brought significant opportunities and challenges.[495] Among the challenges is the proliferation of online hate speech, which has raised urgent questions about how to regulate harmful content in a globalised digital environment.

A stark example is the Christchurch terrorist attack on 15 March 2019, in which the perpetrator was deeply embedded in extremist online communities. Prior to the attack, he published a manifesto espousing white supremacist ideology and then live-streamed the massacre via Facebook Live. His views were shaped and reinforced by online echo chambers that legitimised

[495] Stacy Jo Dixon "Number of social media users worldwide 2010-2021" (April 2020) Statista <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>.

and glorified violence against minority groups. Moreover, the viral spread of his manifesto and video content across social media platforms illustrates how digital ecosystems can not only facilitate radicalisation but also serve as tools for spectacle and recruitment. In this context, online hate speech is not merely abstract or symbolic. It becomes part of a communicative architecture that enables, legitimises and disseminates violent ideology. This example illustrates Waldron's concern that hate speech undermines dignity and social assurance, while also showing Mill's warning that speech crossing into direct harm can justify legal intervention. This raises urgent questions about the adequacy of existing legal frameworks in addressing such threats and holding platforms accountable for the systemic amplification of harm.

This section evaluates global strategies for regulating online hate speech, including censorship, geo-blocking, takedowns, and deplatforming. While these measures can effectively target specific content, they also provoke critical debates about their impact on civil liberties, enforcement challenges, and jurisdictional limitations. By analysing international case studies and legislative frameworks, this chapter seeks to address the research question: How effective is New Zealand's legal framework in combating online hate speech compared to international best practices? This section examines the strengths and limitations of these regulatory measures, focusing on their relevance to New Zealand's efforts to combat online hate speech within its unique legal and social context.

The Christchurch terrorist attack highlighted how social media could be weaponised to livestream and amplify violent content, exposing gaps in content moderation and platform accountability. In response, New Zealand introduced the Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill in 2020. This initiative complements international discussions on regulatory

measures, such as the European Union's frameworks for intermediary liability, by offering insights into platform accountability and the practical challenges of transnational enforcement-key considerations for New Zealand's own framework.

At the international level, the European Union has established comprehensive frameworks to regulate online content. Directive 2000/31/EC, specifically Articles 12–14 of the e-Commerce Directive, outlines the principle of limited liability for service providers, ensuring platforms are not held responsible for third-party content unless they have actual knowledge of its illegality. This principle, which balances platform accountability with the protection of fundamental rights, offers a valuable reference point for New Zealand's efforts to combat online hate speech. From Breyer's proportionality perspective, this balance between liability and rights protection shows how regulation can pursue a legitimate goal without imposing unnecessary burdens. By comparing these approaches, this chapter examines how New Zealand's legal framework can integrate international best practices to address the challenges posed by transnational online hate speech.

5.7.1 Censorship

Censorship has historically served as a tool for regulating speech, initially aimed at suppressing dissent and preserving public order. In the 16th century, European rulers implemented licensing systems to control artistic and scientific expressions perceived as threats to societal morals or political stability.[496] Today, censorship has evolved to address contemporary challenges, including the regulation of harmful online content like hate speech. However, applying

[496] Mette Newth "The Long History of Censorship" (2010) < https://brewminate.com/the-long-history-of-censorship/ >.

censorship to the digital realm introduces complex questions about balancing content regulation with the protection of free expression, particularly in a globalised and highly interconnected internet landscape.

The roots of censorship stretch back to ancient times, with historical examples illustrating its evolving role in regulating speech to preserve societal order. In ancient Greece, Socrates was condemned to death for his perceived moral corruption of youth, underscoring how state-imposed censorship aimed to enforce communal values.[497] Similarly, the Roman Catholic Church in the 16th century established licensing systems to control the publication of texts deemed blasphemous or morally subversive, a practice mirrored by European rulers like King Charles IX of France.[498] These historical instances highlight the enduring tension between suppressing harmful content and safeguarding intellectual and moral freedom-a dilemma that continues to shape debates over the regulation of online hate speech in the modern era.

John Milton's 1644 speech Areopagitica remains a landmark in the history of free speech advocacy, challenging the English Parliament's Licencing Order of 1643, which regulated publishing and censorship. [499] Milton argued that prepublication censorship stifled intellectual growth, undermined public morality, and eroded the public's ability to reason independently by limiting exposure to diverse ideas.[500] This anticipates Mill's later claim that suppressing even offensive speech deprives society of the "collision of ideas" necessary for truth. These principles remain relevant in modern discussions on online hate speech regulation. On the other hand, critics of digital censorship echo Milton's concerns, warning that overregulation could

[497] Newth, above n 496.
[498] Newth, above n 496.
[499] Newth, above n 496.
[500] Kevin R. Davis "John Milton" (6 May 2024) <https://www.mtsu.edu/first-amendment/article/1259/john-milton>.

suppress legitimate discourse, curtail democratic participation, and hinder societal progress.[501] Milton's arguments thus provide a valuable lens for analysing the balance between combating harmful content and protecting free expression in today's digital landscape.

Salman Rushdie, an author who has won the Booker Prize, was almost assassinated by someone who tried to obey the fatwa declared against him, as a result of the publication of his book, The Satanic Verses. A Fatwa is an Islamic legal decision or instruction by an authorised Islamic authority. This specific one was issued by Ayotollah Khomeini, the holy head of Islam.[502] This judgement was made in response to Salman Rushdie's publication of a book that was deemed blasphemous by Muslims. This fatwa, which equated to a death sentence, illustrates the enduring tension between freedom of expression and the perceived need to protect religious or cultural sensitivities. The case underscores how divergent cultural perspectives can frame certain expressions as both artistic liberty and incitement to harm. It also highlights Waldron's point that speech can undermine the assurance of dignity for minority groups, while simultaneously raising Mill's concern that silencing speech may entrench orthodoxy.[503]

Understanding censorship requires differentiating between ex-ante censorship, which prevents the dissemination of content before publication, and ex-post measures, which address harmful material after it has been published.[504] In the digital age, this tension is amplified as the internet enables the rapid and global spread of information with a single click.[505] While early censorship primarily targeted books, modern practices now encompass digital media, including

[501] Ulrike Klinger *Digital Democracy and Public Discourse: Dissonant, Disrupted and Unedited?* (Canadian International Council, Vol 69, No 26, October 2021) at 3.
[502] Jeremy Butterfield "Fatwa" in *Fowler's Concise Dictionary of Modern English* (2016) <https://www-oxfordreference-com.ezproxy.waikato.ac.nz/view/10.1093/acref/9780199666317.001.0001/acref-9780199666317-e-1367>.
[503] Waldron, above n 3, at 34 and 108 & Mill, above n 43, at 80.
[504] Andreas Wiesand "Internet Content Suppression" *Culture and Human Rights: The Wroclaw Commentaries:* (De Gruyter, Berlin, Boston 2016) at 100.
[505] Wiesand, above n 504, at 101.

caricatures, video games, and social media posts.[506] These distinctions are particularly relevant in regulating online hate speech, where the timing and scope of interventions can significantly influence their effectiveness and the protection of civil liberties.

In February 2021, Facebook banned Australian news outlets in response to the proposed News Media Bargaining Code. The law sought to protect media companies by compelling tech giants to pay for news content. While not framed as hate speech regulation, this example illustrates how content control can intersect with market power and public access to information.[507] From Lessig's perspective, this shows how "architecture" (platform design) and "law" interact, with code and regulation together shaping what users can see.[508] While not directly aimed at hate speech, the episode illustrates how states may use regulatory leverage to compel platforms to act in the public interest. This dynamic is relevant to hate speech regulation, where governments face similar challenges in holding powerful intermediaries accountable.

The case of *Google Inc. v. Equustek Solutions Inc*[509], which took place in British Columbia, serves as a significant example of the challenges associated with censorship and regulating online content. In this case, a small tech company, Equustek Solutions (E), filed legal action against its distributor (D), who had rebranded E's products as its own and unlawfully sold them. To further complicate matters, D also misappropriated trade secrets and confidential information belonging to E. Although D initially submitted a defence, it was subsequently abandoned when D fled the jurisdiction.[510]

[506] Wiesand, above n 504, at 99.
[507] Amanda Meade "Australia is making Google and Facebook pay for news: what difference will the code make?" *The Guardian* (9 Dec 2020)
[508] Lessig, above n 84, at 664.
[509] *Google Inc. v. Equustek Solutions Inc.* 2001 [2017] 1 SCR 824.
[510] *Google Inc. v. Equustek Solutions Inc.* 2001 [2017] 1 SCR 824 at [2].

To make matters even more difficult, D also stole trade secrets and sensitive information that belonged to E. D first submitted a statement of defence in response to the allegations. Nevertheless, it was dropped because he fled the scene. [511] The legal questions before the court were: (i) whether Google could be ordered, pending trial, to globally de-index websites operated by the distributor that violated court orders and unlawfully sold E's intellectual property; (ii) whether the Supreme Court of British Columbia had the jurisdiction to grant an injunction with extraterritorial effect; and (iii) if so, whether it was just and equitable to do so.[512] The majority judgment ultimately upheld the global de-indexing order, setting a precedent for the extraterritorial application of court orders in the digital realm.

However, the decision highlighted unresolved challenges, such as (i) avoiding excessive control over illegal content; (ii) the potential utility of geo-location technologies as a less intrusive alternative; and (iii) the jurisdictional complexities inherent in regulating online content with global reach. [513]

In light of this, censorship presents a complex duality in its application. On one hand, it can serve as an effective tool for restricting access to harmful content, including hate speech, which poses a significant threat to societal cohesion and individual safety. On the other hand, the use of censorship to suppress speech raises significant concerns about its potential to infringe on

[511] *Google Inc. v. Equustek Solutions Inc.* 2001 [2017] 1 SCR 824 at [2].
[512] *Google Inc. v. Equustek Solutions Inc.* 2001 [2017] 1 SCR 824 per Per McLachlin C.J. and Abella, Moldaver, Karakatsanis, Wagner, Gascon and Brown JJ.
[513] Dan Svantesson "Supreme Court of Canada challenges the idea of state sovereignty OUPblog" 2017 < https://blog.oup.com/2017/08/supreme-court-canada-state-sovereignty/ >

fundamental human rights, particularly the right to freedom of expression.[514] The balance, as Breyer's proportionality framework suggests, lies in ensuring restrictions are necessary, rationally connected to legitimate aims, and not excessively burdensome. This tension prompts critical reflection on the extent to which censorship can be a viable solution to combating hate speech without encroaching upon democratic principles and individual freedoms.

5.7.2 Geo-blocking

Geo-blocking is a technological measure used to restrict access to online content based on a user's geographical location. This process leverages Internet Protocol (IP) addresses, unique numerical identifiers assigned to devices by Internet Service Providers (ISPs).[515] In the event that a computer sends a request to the server for access to material, the server will be able to determine the appropriate location to deliver the request based on the IP address. [516] When a device sends a request to access online material, the server uses the IP address to determine the user's location and grant or deny access accordingly. For instance, a user in New Zealand attempting to watch an episode of Saturday Night Live on NBC's U.S.-based platform may encounter the message: "This video is not available in your location." Such restrictions exemplify how geo-blocking is employed to control access to region-specific content.

While geo-blocking can serve legitimate purposes, such as compliance with local laws or protecting copyright-protected material, it raises significant concerns. For one, users can bypass these restrictions by utilising tools like Virtual Private Networks (VPNs), which act as

[514] Liza Negriff "The Past, Present, and Future of Freedom of Speech and Expression in the People's Republic of China" (2021) Topical Research Digest: Human Rights In China <https://www.du.edu/korbel/hrhw/researchdigest/china/FreedomSpeechChina.pdf>.
[515] Karl Schaffarczyk "Explainer: what is geoblocking?" (2021) <http://theconversation.com/explainer-what-is-geoblocking-13057>.
[516] Schaffarczyk, above n 515.

intermediaries by routing a user's request through a server in a different location.[517] Similarly, The Onion Router (TOR) network, initially developed to enhance online privacy, has become a popular tool for circumventing geo-blocking.[518] These technologies enable users to access restricted content while masking their locations, which raises challenges for regulators attempting to control harmful online content, including hate speech.

VPNs also play a crucial role in preserving internet freedom, particularly in authoritarian or totalitarian regimes where access to information is heavily censored or restricted. For example, individuals in such countries rely on VPNs to access unbiased news, communicate securely, and evade government surveillance.[519] These legitimate uses of VPNs highlight the importance of balancing privacy and security with the challenges posed by their potential misuse.

In theoretical terms, geo-blocking reflects Lessig's "architecture" modality of regulation, where access is restricted by design.[520] Yet as users bypass these measures, law, market forces, and social norms also shape outcomes, demonstrating the limits of technical controls alone. Critics argue that overreliance on geo-blocking fragments the internet into national enclaves, risking a "splinternet" rather than a coherent global framework. From Breyer's perspective, restrictions must remain proportionate: justified by a proper purpose, necessary to achieve that aim, and not excessively burdensome on lawful expression.[521]

---

[517] Dan Jerker B. Svantesson "Delineating the Reach of Internet Intermediaries' Content Blocking - ccTLD Blocking, Strict Geo-Location Blocking or a Country Lens Approach" 2014 11 SCRIPTed
[518] The Tor Project "The Tor Project" <https://www.torproject.org/about/history/>.
[519] Nikki Mizuguchi "VPN growth highlights global crackdown on internet freedom" (2023) < https://asia.nikkei.com/Business/Technology/VPN-growth-highlights-global-crackdown-on-internet-freedom > and Shahram Akbarzadeh, Amin Naeni, Ihsan Yilmaz, and Galib Bashirov "Cyber Surveillance and Digital Authoritarianism in Iran" (Vol 14, Issue 3, Global Policy, March 2024). <https://www.globalpolicyjournal.com.>The study discusses how Iranians use VPNs to circumvent heavy censorship measures and access global internet resources despite the government's efforts to control cyberspace through surveillance and counterfeit VPNs.
[520] Lessig, above n 84 at 664.
[521] Breyer, above n 122, at 33.

Thus, while geo-blocking can help mitigate the cross-border spread of online hate speech, its effectiveness is undermined by circumvention tools and jurisdictional conflicts. Any adoption of geo-blocking in New Zealand would need to balance security objectives with the protection of civil liberties, ensuring that measures remain proportionate and do not erode the benefits of an open internet.

5.7.3 Strategies for Moderating Online Harm: Web Filter, Takedown, and Deplatforming

In New Zealand, the Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill seeks to modernise the country's legal framework to address harmful online content. Its relevance for this thesis lies in how it demonstrates New Zealand's willingness to adopt urgent measures for content moderation, particularly through the proposal of a government-operated web filter to block offensive material. A central proposal was a government-operated web filter to block offensive material. While the initiative aimed to strengthen platform accountability, critics argued the Bill lacked clarity about its operation and safeguards. [522] Concerns were raised that future governments might expand the scope of the filter, risking overreach and misuse, particularly under more authoritarian leadership. [523]

Currently, New Zealand employs a single web filter, the Digital Child Exploitation Filtering System. Supported by the government, this system is specifically designed to combat the exploitation of children by restricting access to content that promotes sexual abuse or

[522] New Zealand Council for Civil Liberties "Submission on the Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2021"
[523] Andrew Chen "Submission on the Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2021".

exploitative behaviour.[524] While effective in its narrowly defined purpose, the introduction of a broader web filter has provoked fears of unintended consequences, including the potential erosion of civil liberties and over-censorship of legitimate content. From a theoretical perspective, this reflects Breyer's proportionality framework where restrictions may pursue a legitimate aim, but they must also remain necessary and avoid imposing excessive burdens on lawful expression.[525]

Take-down notifications and deplatforming provide additional mechanisms for addressing harmful content online.[526] A takedown notice allows social media platforms to enforce their community guidelines by removing specific content that violates established rules. In more serious cases, platforms may deplatform users entirely, suspending or permanently banning accounts that repeatedly disseminate harmful or inciteful material. From a regulatory perspective, these measures illustrate Lessig's idea of "code as law," where the platform's technical architecture and internal rules function as a form of governance.[527] At the same time, they raise concerns about proportionality in Breyer's sense: while takedowns and bans can be necessary to prevent serious harm, they must not impose excessive burdens on legitimate speech.

The principle of notice and takedown has been widely studied, particularly in relation to combating cyberbullying. Over the past decade, this method has emerged as a core strategy for moderating harmful content while maintaining platform accountability.[528] Yet its application

[524] Rachel Tan "Submission on the Films, Videos, anhd Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2021".
[525] Breyer, above n 122, at 10 & 33.
[526] New Zealand Parliament "Films, Videos and Publications Classification (Urgent...Harm) Amendment Bill - Digest 2626 - New Zealand Parliament" (2021) <https://www.parliament.nz/en/pb/bills-and-laws/bills-digests/document/52PLLaw262611/films-videos-and-publications-classification-urgentharm>.
[527] Lessig, above n 84, at 664.
[528] Brian O'Shea "A New Method to Address Cyberbullying in the United States: The Application of a Notice-and-Takedown Model as a Restriction on Cyberbullying Speech Notes" 2017 69 Fed. Comm. L.J. 121 at 121.

is not without challenges. Scholars point to uneven enforcement across platforms and jurisdictions, which raises concerns about fairness and predictability. More broadly, takedowns highlight the tension between safeguarding users from harm and protecting the freedom of expression. Meiklejohn, advancing a civic republican approach, indicates that restrictions through takedown procedures must remain necessary and balanced, ensuring that interventions do not silence legitimate discourse while addressing harmful content effectively.[529]

A landmark case that highlights the intersection of privacy rights, freedom of expression, and content removal is *Google Spain v González.*[530] This case led to the establishment of Google's Advisory Council, comprising the company's legal division and external communities, to evaluate takedown requests.[531] Mario Costeja González, a Spanish individual who sought to have links to a 1998 newspaper article regarding his unpaid welfare debt removed from Google search results. González argued that the debt had been resolved long ago and that the continued prominence of these links unjustly harmed his reputation.[532] The case sparked global debates over the "right to be forgotten," raising concerns about how privacy rights can coexist with the principles of free expression and transparency online. In theoretical terms, this dispute illustrates Breyer's proportionality framework: the ECJ was required to weigh privacy and dignity interests against the competing public interest in access to information.[533]

[529] Alexander Meiklejohn, *Free Speech and Its Relation to Self-Government* (Harper & Brothers, 1948) at 26.
[530] Case C 131-12 *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González* ECLI:EU:C:2014:317.
[531] Case C 131-12 *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González* ECLI:EU:C:2014:317 at [42].
[532] The Associated Press "Landmark European Court Ruling Could Force Google, Yahoo, Bing To Scrub Online Reputations" (2014) <https://www.cbsnews.com/sanfrancisco/news/landmark-european-court-ruling-could-force-google-yahoo-bing-to-scrub-online-reputations/>.
[533] Stephen Breyer Active Liberty: Interpreting Our Democratic Constitution (The Tanner Lectures on Human Values, Harvard University, 2004) at 10.

A court action that included Costeja González was covered in pieces that were published in the Spanish daily La Vanguardia in the year 1998.[534] The Spanish newspaper, La Vanguardia, which had originally published the article, refused González's request for removal, citing a publication order from Spain's Ministry of Labour and Social Affairs.[535] In 2010, González lodged a formal complaint with Spain's Data Protection Agency against the newspaper, Google Spain, and Google Inc., asserting that search engines should not link to outdated and misleading personal information. [536] While the agency dismissed the complaint against the newspaper, it upheld González's case against Google, ruling that as an operator of a search engine handling personal data, Google bore a responsibility to respect privacy rights and remove harmful links.[537]

Google Spain and Google Inc. both appealed the decision, leading to proceedings before the National High Court of Spain.[538] The court temporarily suspended the case to examine Google's obligations in balancing personal privacy with public access to information. Ultimately, the European Court of Justice ruled that search engines must assess takedown requests to ensure they do not disproportionately infringe on privacy rights. This ruling emphasised the nuanced role of intermediaries in balancing competing rights in the digital age.

The Google Spain case has broader implications for online hate speech regulation, particularly through its conceptual link to deplatforming. Deplatforming, often viewed as a derivative of the "right to be forgotten" within a privacy context, involves the removal of a user's account

[534] Columbia Global Freedom of Expression "Google Spain SL v. Agencia Española de Protección de Datos" <https://globalfreedomofexpression.columbia.edu/cases/google-spain-sl-v-agencia-espanola-de-proteccion-de-datos-aepd/>.
[535] Columbia Global Freedom of Expression, above n 534.
[536] Columbia Global Freedom of Expression, above n 534.
[537] Columbia Global Freedom of Expression, above n 534.
[538] Columbia Global Freedom of Expression, above n 534.

from a social media platform when they violate its guidelines.[539] This strategy extends beyond removing individual pieces of content to addressing the broader harm caused by repeated violations. Platforms may also remove all associated material created by the deplatformed user, ensuring their content cannot continue to propagate harmful narratives. [540] Here, Lessig's "code as law" is visible: technical design (removing or disabling accounts) enforces norms in tandem with law, demonstrating how architecture becomes a form of regulation.[541]

While deplatforming offers a powerful tool for combating online hate speech, it also raises complex questions. These include whether platforms should assume the role of arbiters of acceptable speech and how such measures align with broader commitments to free expression. Furthermore, it highlights the challenge of ensuring that content removed from one platform does not remain accessible through others, such as search engines or online archives. This reflects the broader regulatory difficulty in managing cross-platform consistency and the persistence of harmful content. Cohen's warning resonates here: excessive reliance on private actors to decide the limits of discourse risks chilling legitimate debate.[542]

Deplatforming has been suggested as an effective method for curbing the spread of hate speech online; however, it is not without limitations.[543] An illustrative case is the response to the shootings at a Pittsburgh synagogue and a high school in the United States, where the social media platform Gab became a focal point for controversy.[544] Gab, which operates with fewer

[539] Case C 131-12 *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González* ECLI:EU:C:2014:317.
[540] Richard Rogers "Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media:" (2020) 35(3) European Journal of Communication 213 at 214-215.
[541] Lessig, above n 84, at 664.
[542] Julie E Cohen "Imagining the Networked Society, in Configuring the Networked" (Yale University Press 2012) at 5.
[543] Amin Mekacher, Max Falkenberg, Andrea Baronchelli "The systemic impact of deplatforming on social media" (2023) 2 PNAS Nexus at 6.
[544] Evan Brody, Spencer P. Greenhalgh, and Mehroz Sajjad "Gayservatives on Gab: LGBTQ+ Communities and Far Right Social Media" 8 SM+S at 2.

content moderation policies than mainstream platforms, was used by alt-right groups to spread anti-Semitic and anti-immigrant sentiments. These events highlighted the platform's role in amplifying hate speech and extremist ideologies.[545] In the aftermath of the attacks, cloud infrastructure providers and app stores took decisive action, banning and removing Gab from their platforms. While this deplatforming effort temporarily disrupted Gab's operations and diminished its visibility, it underscored significant challenges. Unlike mainstream platforms such as Facebook or Twitter, which have established content moderation mechanisms, Gab's lax policies made it a haven for harmful speech. [546] This case demonstrates that while deplatforming can disrupt networks of hate, it is not a permanent solution. The persistence of alternative platforms and the potential for re-emergence of harmful content highlight the limitations of this strategy in addressing the root causes of online hate speech. From Waldron's perspective, however, such interventions are justified insofar as they signal to vulnerable communities that dignity and social assurance will not be eroded by unchecked extremist discourse.[547]

Twitter permanently suspended Donald Trump's account following the civil unrest at Capitol Hill on 6 January 2021.[548] [549] Other tech giants, such as Apple and Google, also took similar steps to restrict his presence on their platforms. This marked a significant moment in the debate over the role of private companies in regulating online speech, raising questions about the balance between deplatforming as a tool to combat harmful content and the implications for free speech and democratic discourse. Deplatforming, in this context, may represent a

[545] Jeremy Blackburn and others "Does 'deplatforming' work to curb hate speech and calls for violence? 3 experts in online communications weigh in" (16 January 2021) <https://theconversation.com/does-deplatforming-work-to-curb-hate-speech-and-calls-for-violence-3-experts-in-online-communications-weigh-in-153177>; Lessig, above n 84, at 664.
[546] Abby Vesoulis How Gab Became the Social Media Site Where the Pittsburgh Suspect's Anti-Semitism Thrived *Time Magazine*, (2018).
[547] Waldron, above n 3, at 109.
[548] Khavin Dmitriy and others "U.S. Capitol Riot" *The New York Times* (19 January 2021)
[549] Blackburn and others, above n 545.

paradigm shift, empowering social media platforms to suppress hate speech more effectively while sparking calls for clearer accountability mechanisms to govern these decisions. The episode reflects Breyer's emphasis on proportionality: restrictions may be legitimate if necessary to prevent serious harms, but they must remain carefully bounded to avoid excessive suppression of political speech.[550]

Due to its vast user base, Facebook plays a pivotal role in the regulation of online hate speech. The responsibility placed on social media corporations often surpasses that of legislation or government enforcement, highlighting the unique power of platforms to shape digital discourse. Reports suggest that Facebook removes over 288,000 posts containing hate speech each month, underscoring the scale of the issue and the platform's proactive approach to content moderation.[551]

One key mechanism for regulating hate speech is flagging, where users report content deemed illegal or offensive to the platform. Flagging allows users to raise concerns about violations of Community Guidelines, creating an interactive process that involves users, platforms, algorithms, and the broader social norms governing digital spaces.[552] The Facebook Oversight Board for Facebook and Instagram exemplifies this dynamic. On one hand, its existence highlights gaps in Community Standards policies; on the other, it demonstrates an independent mechanism for appealing platform decisions, aiming to increase accountability and

[550] Breyer, above n 122, at 10 and 27-28.
[551] Wilson Richard Ashby "HATE: Why We Should Resist it with Free Speech, Not Censorship by Nadine Strossen (review)" 2019 41 Human Rights Quarterly 213 at 214.
[552] Kate Crawford and Tarleton Gillespie "What is a flag for? Social media reporting tools and the vocabulary of complaint" (2016) 18 New Media & Society 410 at 411.

transparency.[553] This co-regulatory model resonates with Lessig's idea that law, social norms, and architecture intersect to structure behaviour in digital environments.[554]

A more holistic approach to hate speech regulation has been proposed, emphasizing the need to move beyond legislation to include societal input and systemic balancing of interests.[555] Since the COVID-19 pandemic, there has been a noticeable paradigm shift in how social media platforms moderate content, as they now balance societal interests while weighing the systemic costs of errors.[556] This shift calls for adaptive governance of speech that reflects the evolving digital landscape. Failure to adapt risks undermining the legitimacy of hate speech policies, as effective rule-making requires public trust and acceptance.[557] Public consultation and transparency are essential to ensure that the rules governing hate speech are both effective and aligned with democratic values. This reflects Breyer's insight that proportionality requires continuous recalibration, as well as Waldron's emphasis that policies must visibly secure dignity and reassurance for those targeted by hate.[558]

Cancel Culture

An Australian model and social media influencer, Rawiri Tuapawa[559], widely known as @RaTheKenDoll on platforms such as TikTok, Snapchat, Instagram, and Facebook-rose to prominence due to his advocacy for Maori culture and his physical appearance.[560] With

[553] Evelyn Douek "Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility" (2021) 21 North Carolina Journal of Law & Technology 1 at 7.
[554] Lawrence Lessig "The New Chicago School" (1998) 27 The Journal of Legal Studies 661 at 664.
[555] Evelyn Douek "Governing Online Speech: From "Posts-As-Trumps" To Proportionality And Probability" (2021) 121 Columbia Law Review 759 at 784.
[556] Douek, above n 555, at 759.
[557] Douek, above n 555, at 821.
[558] Breyer, above n 122, at 11-12 and 33; and Waldron, above n 3, at 5.
[559] Rawiri Tuapawa (@RaTheKenDoll) <www.tiktok.com/rathekendoll>
[560] Nicola Gray "Social Media and The Use of 'Clout'" (January 14, 2020) <http://www.theeditgcu.com/arts-culture/social-media-and-the-use-of-clout/>.

approximately 698,000 followers on TikTok alone, his robust social media presence exemplified the concept of “clout,” which refers to influence or popularity on social media platforms.

Tuapawa’s fame peaked in late November 2020, with male followers frequently praising him and engaging with his posts through TikTok’s duet feature. However, in a span of just six days, Tuapawa’s image dramatically shifted. During livestream[561] sessions on TikTok, Tuapawa expressed harmful comments about children, made derogatory statements about women, and voiced hateful remarks toward the LGBTQ+ community. These actions led to widespread backlash, painting him as racist, sexist, homophobic, and ableist.

Instead of issuing an apology, Tuapawa released an eight-part video series attempting to justify his behaviour as a "social experiment." This further inflamed public opinion, as he failed to express remorse for statements advocating that marginalized groups "should not exist and die".[562] Tuapawa framed his remarks as an exercise in free speech, but his actions violated TikTok’s Community Guidelines and possibly its Virtual Items Policy. [563] This prompted speculation that TikTok shadowbanned his account in response to these violations.

Shadowbanning refers to a moderation technique where a user’s content becomes less visible without their explicit knowledge. This practice, reminiscent of early internet forum bans, can significantly limit a user’s reach and engagement. [564] For Tuapawa, the potential shadowban

[561] TikTok “What is TikTok LIVE?” (2020) <https://support.tiktok.com/en/live-gifts-wallet/tiktok-live/what-is-tiktok-live>.
[562] Livestream TikTok @RaTheKenDoll (May 2021) <www.tiktok.com/rathekendoll>
[563] Livestream TikTok, above n 559.
[564] Samantha Cole “Where Did the Concept of 'Shadow Banning' Come From?” (2018) <https://www.vice.com/en/article/a3q744/where-did-shadow-banning-come-from-trump-republicans-shadowbanned>.

had serious repercussions. His primary source of income, sponsorship deals, ceased entirely as sponsors withdrew their support, and his follower count dwindled as his behaviour became widely known.

Public backlash, or what is commonly termed "cancel culture," played a pivotal role in Tuapawa's fall from grace. Social media communities rallied to call for the deactivation of his account. "Cancelling" occurs when public disapproval escalates to the point where a previously popular figure is deplatformed or otherwise loses their influence due to their behaviour. In Tuapawa's case, public sentiment worked in tandem with platform enforcement mechanisms to hold him accountable. The case illustrates both Mill's warning against the tyranny of majority opinion where societal pressure can silence individuals and Waldron's argument that limits on expression are justified where speech undermines the dignity and assurance of minority groups.

It is essential to recognize that the services provided by platforms extend beyond merely hosting applications. These platforms create a sophisticated ecosystem encompassing user licenses, adherence to community standards, and enforcement of consequences for violations. This system establishes a mutually beneficial relationship: users gain access to the platform's services while agreeing to abide by its rules. Platforms, in turn, retain the authority to act against users who breach these standards, ensuring a digital environment conducive to positive engagement. Tuapawa's case exemplifies how these frameworks operate in practice, highlighting the importance of balancing community expectations, individual rights, and platform governance.

5.7.4 Community Guidelines and Oversight Boards

Social media plays a clear role in spreading harmful speech. This raises the question of whether accountability should rest with users, platforms, or both. Major platforms such as Facebook, Instagram, Twitter, and TikTok have adopted Community Guidelines to regulate harmful content. While these guidelines indicate that platforms accept some responsibility, their effectiveness is inconsistent and contested.

This section shows why self-regulation through Community Guidelines remains inconsistent and inadequate. By testing Meta, Twitter/X, and TikTok against New Zealand's legal framework and against free speech theories, the analysis demonstrates why stronger statutory duties may be required. This section argues that Community Guidelines provide only partial protection against hate speech.

5.7.4.1 Meta - Facebook & Instagram

Meta (which owns Facebook and Instagram) has adopted Community Guidelines that prohibit hate speech targeting protected characteristics such as race, ethnicity, religion, sexual orientation, and gender. The rules are designed to balance free expression with the need to protect users from serious harm. Enforcement is carried out mainly through takedowns, content demotion, and account suspensions. [565]

[565] Tan, above n 524.

From the perspective of Mill's harm principle, these rules are justified because they aim to prevent concrete harm rather than mere offence. [566] Waldron's focus on dignity also helps to explain why Meta prohibits speech that undermines people's sense of belonging in society. Similarities exist with New Zealand's HDCA, but enforcement remains inconsistent, raising doubts about voluntary self-regulation.

Breyer's proportionality lens highlights the risk that takedowns may sometimes restrict legitimate expression more than is necessary. Lessig's model also reminds us that "code" shapes enforcement: the design of algorithms and reporting tools determines which speech is removed and which escapes notice.[567] This tension suggests that self-regulation alone may not provide sufficient accountability.

The global scale of Meta means its rules influence democratic participation worldwide, raising concerns about leaving such decisions to private actors. This suggests that relying only on private rules may not be enough to protect participation and equality of voice.[568]

To understand Meta's role in regulating online hate speech, it is crucial to distinguish between self-regulation through its Terms of Service and potential legal obligations under statutory or common law frameworks. The question remains: Does Meta's current approach align with regulatory expectations in jurisdictions like New Zealand, or does it necessitate additional statutory intervention, such as the Harmful Digital Communications Act? Breyer's

[566] John Stuart Mill *On Liberty*, edited by David Bromwich, and George Kateb (Yale University Press, 2003) at 80
[567] Joel R. Reidenberg "Lex Informatica: The Formulation of Information Policy Rules through Technology" , 76 Tex. L. Rev. 553 (1997-1998) 28 at 555-557.
[568] Alexander Meiklejohn, *Free Speech and Its Relation to Self-Government* (Harper & Brothers, 1948) at 26.

proportionality test is useful here.[569] It asks whether restrictions on speech are no more than necessary to protect against harm. Meta's approach sometimes goes further than law, while at other times it falls short, creating uncertainty about the proper balance.

In 2020, Facebook updated its Terms of Service to expand its powers to restrict user access and remove content.[570] This shows the dual nature of platform rules: they can help regulate harmful speech but also protect the company from liability. The change was driven by legal pressure, such as settlements in U.S. case such as, *Fraley v Facebook*[571], rather than voluntary commitment. This reactive pattern suggests that platform rules evolve mainly under legal threat, not proactive responsibility.

Meta's Terms of Service give the company wide rights over user content. In practice, this means the platform decides what is removed or allowed, including speech judged to be hateful. In effect, Meta acts as its own censorship authority. Waldron's dignity-based approach suggests that such discretion is necessary when hateful speech undermines equal membership in society.[572] But Breyer's proportionality framework warns that unchecked platform power can also restrict legitimate expression more than is required.[573]

Instagram, acquired by Meta in 2012, operates under the same Terms of Service as Facebook. This means that content-sharing rights and moderation policies are centralised across the two platforms. The shared framework highlights a regulatory difficulty: one company sets rules

[569] Stephen Breyer Active Liberty: Interpreting Our Democratic Constitution (The Tanner Lectures on Human Values, Harvard University, 2004) at 10.
[570] Stephen Warwick "Facebook is changing its Terms of Service, and users are not happy" (1 September 2020) Windows Central <https://www.windowscentral.com/facebook-changing-its-terms-service-and-users-are-not-happy >
[571] *Fraley v. Facebook, Inc.* 2012 U.S. Dist. 34477 (N.D. Cal. March 13, 2012) at [10].
[572] Waldron, above n 3, at 115.
[573] Waldron, above n 3, at 106.

that affect very different online spaces. [574] From the perspective of Lessig's modalities, the same "code" and contractual terms are applied across platforms, but the social meaning of speech differs between them. This raises questions about whether uniform rules can adequately address hate speech in diverse online communities.

Community Standards and Hate Speech Policies

Meta's Community Standards include several categories, of which "Objectionable Content" is most relevant to hate speech. The policy defines hate speech as a direct attack based on protected characteristics such as race, ethnicity, disability, religion, sexual orientation, or gender identity.[575] This definition is broad and overlaps with New Zealand law, which protects similar categories under the Human Rights Act and the Harmful Digital Communications Act. However, unlike statutory frameworks, enforcement rests on Meta's private discretion.

Because Meta owns both Facebook and Instagram, the same Community Standards apply across both platforms.[576] This centralised approach means that content moderation and hate speech rules are formally aligned. However, algorithmic ranking often privileges engagement over responsibility, making controversial or extremist content more visible. From the perspective of Lessig's modalities, this shows how "code" (ranking systems) can undermine "law" (formal rules). [577] The key dilemma is whether voluntary standards are sufficient to address amplification harms, or whether statutory duties should require platforms to design algorithms with greater responsibility.

[574] Megan Garber "Instagram Was First Called 'Burbn'" (2 July 2014) The Atlantic <https://www.theatlantic.com/technology/archive/2014/07/instagram-used-to-be-called-brbn/373815/>
[575] Facebook "Community Standards - Hate Speech" (May 2023) <https://www.facebook.com/communitystandards/hate_speech>.
[576] Nick Clegg "Facebook Does Not Benefit from Hate - About Facebook" (2020) <https://about.fb.com/news/2020/07/facebook-does-not-benefit-from-hate/>.
[577] Lawrence Lessig "The New Chicago School" (1998) 27 The Journal of Legal Studies 661 at 664.

Algorithms shape what users see by prioritising posts that drive engagement, even when this means amplifying controversial or hateful material. When left unchecked, such systems can create echo chambers that reinforce harmful narratives. Meta defines hate speech as a direct attack based on characteristics such as race, disability, religion, gender identity, or serious disease[578]. This broad definition mirrors statutory protections but is enforced only at Meta's discretion.

Facebook's policy of permitting hate speech under the pretext of 'raising awareness' presents a regulatory challenge.[579] While aligned with U.S. free speech norms, this loophole may allow harmful content to persist. Waldron's dignity-based theory shows why such exceptions are dangerous: even when labelled as "awareness," targeted groups still experience harm to their social standing.[580] In contrast, Meiklejohn's view could justify the exception in the name of democratic debate.[581] This tension raises the central regulatory question: should Meta be held to a statutory duty of care, or is its self-regulation and Oversight Board an adequate safeguard? Meta has reported large volumes of takedowns, but the scale of online content means harmful material often circulates faster than it can be removed. [582]

Context is a persistent challenge for Meta. For example, the term "dyke" may be flagged as hate speech in some settings but allowed in others where it is reclaimed as identity.[583] This

[578] Facebook, above n 575.
[579] Facebook, "Facebook Community Standards" (2025) <https://transparency.meta.com/en-gb/policies/community-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2F > .
[580] Waldron, above n 3, at 34 and 108.
[581] Alexander Meiklejohn, *Free Speech and Its Relation to Self-Government* (Harper & Brothers, 1948) at 26.
[582] Richard Allan "Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community? - About Facebook" (2017) <https://about.fb.com/news/2017/06/hard-questions-hate-speech/>.
[583] Facebook, above n 575 and Richard Allan "Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community? - About Facebook" (2017) <https://about.fb.com/news/2017/06/hard-questions-hate-speech/>.

illustrates the difficulty of balancing protection from harm with freedom of expression. Breyer's proportionality model helps frame this: regulation must weigh the harm to dignity against the cost of restricting lawful speech.

To improve accountability, Meta established the Oversight Board in 2020 as an independent appeals mechanism. The Board reviews high-profile content disputes and is designed to provide a degree of transparency beyond internal moderation.[584] A notable early case involved the use of a derogatory term against Azerbaijanis. The Oversight Board upheld Meta's removal of the post, reasoning that the slur caused harm to dignity and violated the Community Standards. [585] . The Board upheld Meta's removal, reasoning that the slur violated Community Standards and caused harm to dignity, while also drawing on Article 19 of the ICCPR.

The Oversight Board therefore represents an attempt at independent self-regulation. It enhances transparency but its powers remain limited: it cannot compel wider industry change or enforce rulings beyond Meta's platforms.

According to the Oversight Board's Q1 2023 Transparency Report, users primarily submitted appeals to restore content which Meta removed for violating its policies on Violence and Incitement (37%), Hate Speech (19%), Adult Nudity and Sexual Activity (14%), and Bullying and Harassment (11%).[586] These figures highlight the prevalence of harmful content on

[584] Allan, above n 583.
[585] Case decision 2021-003-FB-UA 2021, Oversight Board Upholds Facebook Decision In Armenians In Azerbaijan Case (2021) <https://www.oversightboard.com/news/436612660860568-oversight-board-upholds-facebook-decision-case-2020-003-fb-ua/ >
[586] Oversight Board *Q1 2023 Transparency Report* (March 2023) <https://www.oversightboard.com/news/1008878700278435-q1-2023-transparency-report-board-publishes-new-data-on-the-impact-of-its-recommendations/>

Facebook and the platform's struggle to balance free expression with content moderation. Additionally, the introduction of a user appeals process suggests an ongoing effort to improve transparency and enhance content regulation mechanisms. While the Oversight Board serves as an important check on Meta's power, the evolving landscape of online hate speech continues to test the effectiveness of its interventions.

The data indicates that online hate speech is a prevalent issue on Facebook, making up a significant proportion of the cases reviewed by the Oversight Board. Users also encounter a wide range of harmful content, including violence, adult nudity, and support for dangerous individuals or organisations. The introduction of a user appeals process for content removal shows an effort to improve moderation practices and address concerns about objectionable material. Taken together, these insights highlight the complex landscape of content regulation on social media and underline the need for robust mechanisms to address harmful content effectively.

5.7.4.2 Twitter/X[587]

In July 2023, Twitter underwent a significant rebranding, adopting the name "X" as part of Elon Musk's vision to transform the platform into an "everything app."[588] Originally launched in 2006 as a microblogging site, Twitter quickly evolved into a central platform for political communication, activism, and public debate.[589] Its global user base and real-time nature give it unique influence in shaping public discourse. For example, U.S. President Donald Trump's prolific use of Twitter demonstrated its potential to shape political narratives: his references to

[587] Twitter was rebranded as 'X' under Elon Musk's ownership. However, this section will continue to refer to the platform as Twitter for consistency and historical accuracy.
[588] Jordan Valinsky "Elon Musk rebrands Twitter as X" (July 2023) CNN < https://edition.cnn.com/2023/07/24/tech/twitter-rebrands-x-elon-musk-hnk-intl/index>
[589] Katrin Weller and others *Twitter and Society* (Peter Lang, New York, 2013) at 29.

COVID-19 as the "Chinese virus" and "kung flu" were widely criticized for promoting xenophobia and misinformation.[590] This shows how algorithmic amplification can spread harmful speech with real social consequences, connecting to Waldron's concern with dignity and belonging.[591]

Twitter's Hate Speech Policies and Enforcement Mechanisms

Twitter's platform rules are designed to facilitate public conversation while prohibiting content related to violence, harassment, and behaviour that discourages participation. [592] Its policies have shifted significantly after Musk's takeover, with the dismantling of bodies like the Trust & Safety Council raising questions about the consistency of enforcement.

In response to criticism, Twitter/X expanded its rules to prohibit "dehumanising speech against religious groups" and later extended this to other protected characteristics. Under its Hateful Conduct Policy, Twitter/X prohibits attacks based on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religion, age, disability, or serious disease. [593] Mill's harm principle offers justification here: prohibiting dehumanising speech aims to prevent tangible harm, not just offence.[594] At the same time, Breyer's proportionality test reminds us that enforcement must balance this against the risk of curbing lawful democratic expression.[595]

[590] Global Times "Trump's racist words spark hatred, fuel global xenophobia " (2020) <https://www.globaltimes.cn/content/1183207.shtml>.
[591] Waldron, above n 3, at 115.
[592] Twitter "Hateful Conduct Policy" (2021) <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>.
[593] Sarah Perez "Twitter expands hateful conduct rules to ban dehumanizing speech around age, disability and now, disease – TechCrunch" (2020) <https://social.techcrunch.com/2020/03/05/twitter-bans-hate-speech-around-age-disability-and-in-the-wake-of-the-coronavirus-outbreak-disease/>. and Twitter, above n 591.
[594] Mill, above n 43, at 80.
[595] Breyer, above n 122, at 10.

Action is taken when violent threats are made against a class of people, and the result of this may be permanent suspension of the violator's account. Twitter/X claims a zero-tolerance approach to violent threats, with sanctions ranging from labelling and visibility reduction to suspension or removal. In practice, enforcement has been criticised as inconsistent and politically selective.[596]

Enforcement Actions for Hate Speech Violations

Twitter/X employs a tiered system: content-level (e.g. labels, removals), account-level (e.g. suspensions, read-only mode), and structural measures (e.g. authentication).[597] In theory, this balances deterrence with proportionality, but in practice it relies heavily on user reporting and reactive moderation.

Twitter/X applies enforcement at multiple levels: individual content (labels, visibility limits, or removal), direct messages (blocking further contact), and accounts (suspension or permanent bans).[598] While this tiered system appears comprehensive, it relies heavily on user reporting and reactive moderation. Lessig's model suggests that such reliance on "law" and "norms" (policies plus user vigilance) is weak unless backed by "code" that proactively shapes behaviour. The inconsistency of enforcement, particularly for high-profile accounts, shows why law may need to step in where platform rules fall short.

One critical weakness is its reliance on reactive rather than proactive moderation. Lessig's model helps explain why: the platform's "code" and algorithmic incentives prioritise

[596] Twitter, above n 592.
[597] Twitter, above n 592.
[598] Twitter "Staying safe on Twitter and sensitive content" (2021) <https://help.twitter.com/en/forms/safety-and-sensitive-content/abuse>.

engagement, which may amplify hateful material faster than rules can suppress it. High-profile accounts often receive leniency, while ordinary users face stricter enforcement. [599] This inconsistency erodes legitimacy and highlights Murray's point that regulation is most effective when aligned with technical infrastructure rather than imposed after the fact.

5.7.4.3 TikTok

Launched in China in September 2016, TikTok expanded globally the following year, quickly becoming one of the fastest-growing social media platforms. Owned by ByteDance, TikTok is distinctive for its AI-driven recommendation system, which determines visibility of content and can amplify harmful as well as beneficial speech.[600]

TikTok's rapid growth across diverse jurisdictions has brought regulatory scrutiny, especially over its handling of hate speech, misinformation, and extremist content.[601] TikTok's Community Guidelines were updated in March 2023 to reinforce policies on self-harm, dangerous acts, harassment, and violence while introducing Community Principles aimed at balancing freedom of expression with harm prevention.[602] These rules define hateful behaviour as content that "attacks, threatens, dehumanises, or demeans" people based on protected attributes, now including "tribe" and immigration status.[603] From a Waldron-ian perspective, TikTok's inclusion of immigration status shows recognition that dignity and equal membership

[599] Giulio Corsi "Evaluating Twitter's algorithmic amplification of low-credibility content: an observational study" (2024) 13 EPJ Data Science <https://doi.org/10.1140/epjds/s13688-024-00456-3>
[600] Iqbal, above n 210.
[601] Madhav Chanchani "India clocks over 5.5 billion hours on TikTok in 2019" (2020) Diligent < https://timesofindia.indiatimes.com/business/india-business/india-clocks-over-5-5-billion-hours-on-tiktok-in-2019/articleshow/73787861.cms>
[602] Sarah Perez "TikTok expands Community Guidelines, rolls out new 'well-being' features –" (2020) TechCrunch < https://techcrunch.com/2020/12/15/tiktok-expands-community-guidelines-rolls-out-new-well-being-focused-features/>.
[603] Julie de Bailliencourt" Helping creators understand our rules with refreshed Community Guidelines" (Mar 2023) < https://newsroom.tiktok.com/en-us/community-guidelines-update>

in society are threatened by rising hostility towards migrants.[604] At the same time, Lessig's modalities highlight that the "code" of TikTok's algorithm often overrides the "law" of its rules by pushing controversial material for engagement.

In addition to guarding its rules, the platform also moderates its hashtags by removing or redirecting controversial ones. For instance, during the 2020 U.S. elections, TikTok blocked searches for hashtags like "#RiggedElections" and "#SharpieGate." Users searching for these terms would encounter a blank page with zero results and a message stating that "the phrase may be associated with behaviour or content that violates our guidelines."[605] While this demonstrates a degree of platform regulation, it raises questions about content removal appeals and whether TikTok offers a redress mechanism similar to Facebook and Instagram's Oversight Board.

TikTok's Community Guidelines encompass various components outlining user conduct on the platform, addressing issues such as violent extremism, adult nudity and sexual activities, minor safety, self-harm, suicide, harassment and bullying, and hateful behaviour. [606] The platform defines hate speech or behaviour as content that "attacks, threatens, dehumanizes, or demeans an individual or group based on protected attributes."[607]. These protected attributes include race, ethnicity, national origin, religion, caste, sexual orientation, sex, gender, gender identity, serious disease, disability, and immigration status.[608] Notably, TikTok's inclusion of

[604] Waldron, above n 3, at 85
[605] Sarah Perez "TikTok takes down some hashtags related to election misinformation, ignores others" (2020) TechCrunch < https://techcrunch.com/2020/11/05/tiktok-takes-down-some-hashtags-related-to-election-misinformation-leaves-others/>.
[606] TikTok "Safety and Civility" (2024) < https://www.tiktok.com/community-guidelines/en/safety-civility>.
[607] TikTok, above n 606.
[608] TikTok "Countering Hate on TikTok" (2023) < https://www.tiktok.com/safety/en/countering-hate/>

immigration status as a protected attribute indicates an awareness of rising hate related to migrant issues.[609]

TikTok also defines slurs as derogatory terms specifically intended to disparage individuals based on their protected attributes.[610] In addition, TikTok outlines hateful ideology as content that "demonstrates clear hostility toward people because of their protected attributes".[611] The Community Guidelines explicitly state that TikTok will not hesitate to remove or take down content that violates its policies, reinforcing its commitment to countering hate speech. This includes any hateful content that is posted, uploaded, streamed, or shared.[612]

Social media networks fundamentally operate within the cyber world, a realm governed by codes and algorithms. Understanding how to regulate online hate speech requires an awareness of this digital ecosystem. At the core of social media regulation lies algorithmic science, which dictates content visibility, moderation, and user interactions.[613]

Algorithmic failures are well documented. Reports show TikTok has suppressed LGBTQI, disabled, and "unattractive" users to maintain a certain aspirational image.[614] This discriminatory filtering reflects Breyer's concern with proportionality: protecting one group (e.g. by limiting political or controversial speech) may unnecessarily infringe on lawful expression by others.

[609] Ciarán O'Connor *Hatescape: An In-Depth Analysis of Extremism and Hate Speech on TikTok* (Institute for Strategic Dialogue, 2021) at 18.
[610] TikTok, above n 608.
[611] TikTok, above n 608.
[612] TikTok, above n 608.
[613] Terevinto and others, above n 218, at 47.
[614] Köver and Reuter, above n 230.

In New Zealand, no statutory framework currently requires algorithmic transparency. By contrast, the EU's Digital Services Act and the UK's Online Safety Act 2023 impose stronger obligations, including proactive detection of harmful content. This gap raises the question: should New Zealand adopt a statutory duty of care to ensure fairness and accountability in algorithmic design, or continue to rely on voluntary guidelines?

TikTok has also been criticised for slow removal of harmful viral content. For example, an anti-Semitic Holocaust song reached nearly half a million views before deletion eight hours later.[615] The song was viewed nearly 500,000 times before TikTok removed it eight hours after the initial upload. While this response may seem swift, concerns remain that even short-lived viral content can cause lasting harm, particularly among younger audiences who are more susceptible to extreme racist narratives.[616] Even short delays matter given the speed of viral spread. Here, Murray's theory of dynamic regulation is relevant: effective solutions may require hybrid cooperation between law, platforms, and civil society rather than reliance on self-regulation alone.[617]

The Role of Social Media in Regulating Hate Speech: Gaps and Challenges

In New Zealand, no statutory duty currently mandates social media platforms to proactively monitor and remove extremist content. Instead, moderation is left to voluntary company policies, such as the Community Guidelines of Facebook, X, or TikTok. By contrast, the EU's Digital Services Act and the UK's Online Safety Act 2023 impose strict timeframes for takedown of illegal hate speech, with fines for non-compliance. The Christchurch Call to

[615] Crystal Wu "TikTok algorithm promoted anti-Semitic death camp video" <https://www.newshub.co.nz/home/world/2020/07/tiktok-algorithm-promoted-anti-semitic-death-camp-video.html?utm_source=dlvr.it&utm_medium=twitter>
[616] Wu, above n 615.
[617] Andrew Murray *The Regulation of Cyberspace* (Routledge-Cavendish, Abingdon, 2007) at 240.

Action (2019), launched after the Christchurch terrorist attacks, reflects New Zealand's leadership in this area but remains non-binding.[618] This gap illustrates Murray's point: without binding duties, "dynamic regulation" depends too much on goodwill and industry discretion.[619]

Given that three-quarters of New Zealand's population actively use social media (primarily via mobile devices) the rapid spread of extremist content remains a significant concern.[620] This highlights a regulatory gap: should New Zealand adopt statutory duties similar to the UK and EU, or continue relying on voluntary self-regulation? The answer to this question is central to evaluating whether the current framework is adequate or whether reform is required.

The Power and Risks of Social Media in Shaping Public Discourse

Social media has become a powerful tool in shaping public discourse, enabling rapid mobilisation and amplifying voices. The Black Lives Matter movement illustrates how online platforms can drive global activism across the United States, Europe, and beyond.[621] Yet the same mechanisms that empower social justice movements also facilitate the spread of hate speech and extremist content, raising questions about whether voluntary moderation is sufficient to protect democratic debate.[622]

At the same time, the same digital tools that enable mobilisation also facilitate the spread of hate speech and misinformation. Online platforms allow like-minded individuals to connect rapidly, creating echo chambers where harmful views can be amplified.[623] The anonymity of

[618] Global Internet Forum to Counter Terrorism, above n 145.
[619] Murray, above n 618, at 240.
[620] Christopher Hughes "Most active social media networks New Zealand 2018" (2019) <https://www.statista.com/statistics/681840/new-zealand-most-popular-social-media-networks/>.
[621] Henderson, above n 238.
[622] Maqbool, above n 236.
[623] Enikolopov and others, above n 237 at 1482.

social media lowers social and psychological barriers, encouraging disinhibited behaviour and making extreme speech more common than in offline settings. Research shows that online interactions reduce the personal risks of expressing such views, allowing both productive and harmful discourse to flourish.[624]

Given the growing harm caused by online hate speech and extremism, it is necessary to ask whether the legal framework assigns accountability to end-users, platforms, or both.[625] Current moderation measures remain largely reactive, relying on platform self-regulation rather than systemic reform. This reliance has proven insufficient to curb the persistence and virality of harmful content.

Lawrence Lessig's regulation theory highlights that legal intervention alone is insufficient to regulate online spaces. Instead, effective governance must consider four key forces: law, social norms, market pressures, and architecture (code).[626] From this perspective, online hate speech is not merely a legal problem, but also an issue of platform design and user behaviour. The structural architecture of social media, its algorithms, anonymity, and amplification mechanisms, can facilitate the spread of harmful content, making it a systemic issue beyond just legal enforcement.

The Christchurch terrorist attacks in 2019 illustrate these challenges. Social media was central to both the planning and live-streaming of the violence, with the footage rapidly shared across platforms. The case highlights the need for a multi-layered regulatory approach. Murray's theory of dynamic regulation supports this, arguing that effective governance depends on

[624] Enikolopov and others, above n 237 at 1482.
[625] Binny and others, above n 240, at 372.
[626] Lessig, above n 84 at 664.

collaboration between lawmakers, technology companies, and civil society. A hybrid model that combines legal mandates, platform accountability, and algorithmic oversight is therefore more likely to address online harms than reliance on either government regulation or corporate self-regulation alone.

Legal Jurists Barbara Perry and Patrik Olsson prescribes four mechanisms to curtail online hate speech:

> (1) "filtering, which allows the prevention of specified content from accessibility;
>
> (2) a development of monitoring organisations;
>
> (3) hate speech hotlines where users report incidents which is then sent to law enforcement; and
>
> (4) self-regulation by implementing codes of conduct for their members"[627].

Perry and Olsson outline four mechanisms to address online hate speech: filtering content, developing monitoring bodies, reporting hotlines, and self-regulation through codes of conduct. These illustrate the range of strategies available, from proactive content moderation to legal enforcement.[628] Their effectiveness, however, depends on consistent implementation, which remains limited in New Zealand's largely voluntary framework.

The role of social media in shaping public discourse remains contested. The January 6, 2021 Capitol riots showed how platforms can amplify extremism, leading to the unprecedented

[627] Barbara Perry & Patrik Olsson "Cyberhate: The globalization of hate" (2009) 18 Information & Communications Technology Law 185 at 191.
[628] Perry and Olsson, above n 627; Jialun 'Aaron' Jiang and others "Characterizing Community Guidelines on Social Media Platforms" (paper presented to Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing, Virtual Event, USA, 2020) at 288.

decision to permanently ban a sitting U.S. president from Twitter.[629] While European Commission President Ursula von der Leyen accepted the justification, she warned that leaving such decisions solely to private companies raises serious democratic concerns. As she noted, "We cannot accept that decisions that have a far-reaching impact on our democracy are taken by computer programs alone."[630] This highlights a key tension: Meiklejohn's democratic theory values open participation, yet Waldron's concern for dignity warns against unchecked harm. The episode strengthens calls for legal oversight to ensure that decisions about online speech are not left entirely to corporate discretion.

Taken together, these case studies confirm that platform guidelines provide partial but inconsistent protection against online hate speech. They rely heavily on company discretion, are vulnerable to ownership or policy shifts, and fail to address algorithmic amplification. These weaknesses explain why New Zealand cannot rely on voluntary guidelines alone, and why statutory duties of care or refined Safe Harbour provisions are necessary.

As New Zealand grapples with the challenges of online hate speech, a central question is whether to follow international regulatory models and impose stricter obligations on platforms, or to continue relying on self-regulation. The Christchurch Call to Action (2019) illustrates New Zealand's leadership in fostering cooperation, but its non-binding nature limits its effectiveness. This highlights the need to ask whether the Harmful Digital Communications Act, as the cornerstone of New Zealand's framework, is adequate to address amplification and virality risks on platforms such as TikTok. Lessig's modalities remind us that law alone may not be sufficient; architecture (algorithms) and market incentives also shape how harmful

[629] Ian Wishart "EU Chief Takes Aim at Internet Giants Over Freedom of Speech" (2021) <https://www.bloomberg.com/news/articles/2021-01-26/eu-chief-takes-aim-at-internet-giants-over-freedom-of-speech>.
[630] Wishart, above n 629.

speech spreads. The issue, then, is whether statutory reform should extend beyond individual liability to impose clearer duties of care on platforms themselves.

When considered together, Meta, Twitter/X, and TikTok illustrate the strengths and limits of Community Guidelines. While each recognises hate speech as harmful, enforcement is inconsistent, vulnerable to ownership or design choices, and limited by voluntary compliance. Algorithmic amplification in particular shows how 'code' can undermine stated rules. These weaknesses suggest that New Zealand cannot rely solely on self-regulation and should consider statutory duties of care or reforms to Safe Harbour provisions.

5.8 Intermediary Liability and Safe Harbour Provisions

Internet intermediaries are entities that facilitate online interactions and access to digital content.[631] These intermediaries can be broadly categorised into two types: conduits and hosts. Conduits provide the technical infrastructure of the internet, such as Internet Service Providers (ISPs), while hosts include platforms like Facebook and Twitter that enable the creation, storage, and dissemination of content.[632]

Laidlaw introduces the concept of Authority Gatekeepers to describe intermediaries on the internet, which include social networking platforms such as Facebook.[633] These intermediaries wield significant influence in regulating online interactions and content dissemination. Additionally, Laidlaw proposes the concept of micro-gatekeepers-individuals responsible for

[631] Association for Progressive Communications "Frequently asked questions on internet intermediary liability | Association for Progressive Communications" (2020) <https://www.apc.org/en/pubs/apc%E2%80%99s-frequently-asked-questions-internet-intermed>.
[632] Association for Progressive Communications, above n 631.
[633] Emily B. Laidlaw "A Framework for Identifying Internet Information Gatekeepers" (2010) 24 International Review of law, Computers & Technology 263 at 264.

administering specific pages or groups on platforms like Facebook, who also function as hosts within their domain.[634]

ISPs argue their role is technical and should not attract liability. [635] Despite these assertions, ISPs can still be held accountable for information or material under specific circumstances. Generally, liability arises based on three primary factors: knowledge of the content, control over its dissemination, or direct financial benefit derived from it. [636] Among these, knowledge serves as the most critical determinant of an ISP's accountability, as it directly links the intermediary to the nature and impact of the content being transmitted.

There are essentially two policy justifications that are enforced on social media networks. The first justification holds secondary offenders accountable within the realms of ethics and personal civil liability, emphasizing moral responsibility for harm caused by unregulated content.[637] The second justification applies principles of law and economics to determine why certain secondary actors, such as ISPs, may be appropriate targets for loss-shifting.[638] This framework not only underscores the rationale for intermediary accountability but also highlights the interplay between economic incentives and ethical responsibilities.

There are two key categories of agreements that define the relationship between Internet Service Providers (ISPs) and end users. The first involves a contractual agreement in which the ISP directly provides network services to the end user, thereby assuming a primary role in

[634] Laidlaw, above n 633 at 264.
[635] Jaani Riordan, *The Liability of Internet Intermediaries* (Oxford University Press, Oxford, 2016)
[636] E. Eugene Clark *Cyber law in Australia* (Kluwer Law International, Alphen aan den Rijn, The Netherlands, 2010)318 at 314.
[637] Giancarlo Frosio *Oxford Handbook of Online Intermediary Liability* (Oxford University Press, Oxford, 2020) at 228.
[638] Association for Progressive Communications, above n 631.

facilitating internet access and associated activities.[639] This role may place the ISP in a predicament, particularly when its services are directly tied to harmful conduct or content dissemination.[640] In the second category, ISPs assume a secondary role, functioning as intermediaries rather than primary participants. This occurs, for example, when a user uploads malicious or harmful content onto the internet using the ISP's infrastructure. [641] In such cases, the ISP is indirectly linked to the activity but may still face scrutiny over its responsibility to moderate or prevent the dissemination of such content.

Legal jurist Jaani Riordan observes that internet intermediaries represent a diverse and complex collection of entities that do not conform to uniform norms or guidelines.[642]

There are essentially three types of legal liabilities applicable as follows:

1. Primary liability – This refers to the direct legal responsibility of a party for their actions. In New Zealand, under the Harmful Digital Communications Act 2015 (HDCA), individuals who post harmful content online can be held primarily liable if the content causes serious emotional distress to another person. This Act empowers affected parties to bring complaints directly against the individuals responsible for posting harmful content, thereby demonstrating a clear application of primary liability. For present purposes, primary liability illustrates the limits of New Zealand's current approach: the burden rests almost entirely on end-users, with little attention to the structural role of platforms in amplifying or spreading harmful material.

[639] Clark, above n 636, at 316.
[640] Clark, above n 636, at 316.
[641] Clark, above n 636, at 316.
[642] Riordan, above n 635, at 686.

2. Secondary liability – also identified as *accessory liability,* this arises when a party is deemed legally responsible for contributing to or facilitating the wrongdoing of another.[643] In cases of copyright infringement in New Zealand, internet intermediaries could potentially be held secondarily liable for hosting or facilitating access to infringing material. The Copyright Act 1994 was amended in 2011 to include provisions for intermediary liability. While intermediaries are not directly liable for content uploaded by users, they may be required to act on takedown notices to avoid secondary liability for facilitating infringement.[644] This example shows how New Zealand law already accepts the principle that intermediaries may bear responsibility when they facilitate harm but only in narrow fields like copyright. The question, then, is whether a similar logic could extend to hate speech, given its societal harms.

3. Injunctive liability- This form of liability obligates compliance with specific terms rather than requiring parties to pay monetary damages. Injunctive relief is often employed to prevent ongoing harm or enforce adherence to regulatory requirements.[645] Under New Zealand's HDCA, the District Court can issue orders requiring internet intermediaries to remove harmful content. These orders act as a form of injunctive relief, ensuring compliance with the law and providing immediate protection to victims of harmful online behaviour. In practice, injunctive relief is valuable but reactive. It does not prevent harm before it occurs; instead, it relies on victims pursuing court orders. This highlights a central limitation of New Zealand's current Safe Harbour model: protection depends on post-hoc remedies, not proactive duties of care.

Understanding the origins of these liabilities is essential to framing the regulatory landscape for internet intermediaries. Drawing on the approaches of European jurisdictions, the sources of these obligations can be categorised into three broad areas: Domestic Legal Frameworks -

[643] Riordan, above n 635, at 13.
[644] Riordan, above n 635, at 13.
[645] Riordan, above n 635, at 14.

In New Zealand, the HDCA and Copyright Act serve as key examples of domestic laws that impose primary and secondary liabilities on intermediaries while also offering avenues for injunctive relief; European Union Law - While not directly applicable to New Zealand, EU directives such as the e-Commerce Directive (2000/31/EC) provide a comparative framework for intermediary liability, balancing limited liability with obligations to act upon notice of illegal content; and Human Rights Obligations - New Zealand's adherence to international treaties like the International Covenant on Civil and Political Rights (ICCPR) informs its domestic approach to balancing intermediary liability with the protection of fundamental rights, such as freedom of expression.[646]

By examining these sources and examples, New Zealand can better understand the application of liabilities and adapt its regulatory framework to address the complexities of the modern digital landscape.

As mentioned in Chapter 4, the European method of controlling offensive content follows a horizontal approach, specifically in relation to the E-Commerce Directive. [647] This approach acts as a comprehensive umbrella, spanning all platforms and addressing all forms of illegal or harmful content without being tied to specific issues. Moreover, the EU's regulatory framework incorporates a vertical strategy, which outlines precise requirements for targeted areas such as copyright, child safety, personal data security, and anti-counterfeiting measures. [648] These vertical regulations address specific challenges in a focused manner, complementing

[646] Riordan, above n 635, at 15.
[647] Kyriakos Fountoukakos, Susan Black and Kristien Geeurickx "The EU Commission Adopts New Horizontal Cooperation Guidelines And Publishes the R&D And Specialization Block Exemption Regulations Which Introduce A New Era For Horizontal Cooperation In Line With The Court Of Justice's And The Commission's Decisional Practices" (1 June 2023) <https://www.concurrences.com/en/bulletin/news-issues/june-2023/the-eu-commission-adopts-new-horizontal-cooperation-guidelines-and-publishes-113508>.
[648] Riordan, above n 635, at 11.

the broader principles of horizontal regulation.[649] This layered approach ensures both flexibility and specificity in managing harmful online content.

In New Zealand, the Harmful Digital Communications Act 2015 (HDCA) represents a central regulatory framework; establishes principles for addressing harmful digital communications, such as those targeting an individual's race, gender, or sexual orientation, thereby creating a cascading regulatory framework from general laws to more specific applications.

The EU's mix of horizontal and vertical rules, and New Zealand's more fragmented approach through the HDCA and related laws, both offer lessons for hate speech regulation. The HDCA is important, but it may not match the level of detail and coherence of the EU model. This raises the key question for this chapter: whether New Zealand's framework is strong enough to address hate speech on social media, or whether further duties on platforms are needed.

Although the landscape of social media has evolved significantly over the past decade, the law was not originally designed to address the unique harms caused by these platforms. Social media has become both a means of communication and a form of escapism, particularly during global crises like pandemics. [650] However, despite the benefits these platforms provide for virtual engagement, there is a recurring pattern of harassment and bullying through hateful speech, causing considerable societal harm.[651]

[649] Micova and de Streel, above n 469, at 36.
[650] Rachel Sue Yin Tan "Disabling access to illegal online content by way of takedowns" [2021] NZLJ 341 at 3.
[651] Nikki MacDonald "Online harassment: the insidious face on an inescapable harm" (11 March, 2019) Stuff <https://www.stuff.co.nz/national/crime/110956646/online-harassment-the-insidious-face-on-an-inescapable-harm>.

Comparative law highlights different balances. In the European Union, the earlier e-Commerce Directive offered a similar safe harbour, but the Digital Services Act has moved toward duties of care, risk assessments, and stronger oversight. In New Zealand, the HDCA provides a central framework, but its fragmented approach may not fully address platform-level risks.

Theory from Chapter 2 helps to understand limits of liability rules. Lessig's model shows that law is only one regulatory force; code and market incentives also shape behaviour online.[652] Murray suggests regulation is more effective when it adapts to how platforms are built and operated.[653] These insights indicate that safe harbours should be paired with measures that address platform design and incentives.

European regulation has often combined a general, horizontal framework with targeted, vertical rules. New Zealand's approach is more incremental, with the HDCA sitting alongside other statutes. This provides important tools but may not fully address platform-level risks. The direction for this next section is therefore to consider modest recalibration of safe harbours and stronger expectations on platforms, while leaving final design details to the conclusion and to the Community Guidelines analysis in section 6.3.

5.8.1 Introduction to Safe Harbour provisions

As outlined in the previous section on duty of care, a central legal question is whether online platforms should be treated as neutral intermediaries or as entities bearing heightened responsibility for the content they facilitate. Safe Harbour provisions sit at the heart of this debate. They define the circumstances in which platforms may be exempt from liability for

[652] Lawrence Lessig "The New Chicago School" (1998) 27 The Journal of Legal Studies 661 at 664.
[653] Andrew Murray *The Regulation of Cyberspace* (Routledge-Cavendish, Abingdon, 2007) at 240.

user-generated content, shaping the balance between innovation, free expression, and protection from harm.

The term Safe Harbour refers to legal provisions that protect online intermediaries, such as social media platforms, search engines, and internet service providers (ISPs), from liability for third-party content they host, provided they meet specific statutory requirements. These provisions were originally intended to encourage the growth of digital platforms by limiting excessive legal burdens. At the same time, they impose conditional obligations, requiring platforms to act on notice of illegal or harmful content.[654] This balance between innovation and accountability lies at the core of debates about whether Safe Harbour regimes remain adequate in the age of algorithmic amplification and online hate speech. From Lessig's perspective, Safe Harbour regimes reflect the force of law as a regulatory tool, but they often fail to engage with the role of architecture (algorithms and platform design) in shaping user behaviour. This tension raises the question of whether a notice-and-takedown model, designed for a pre-algorithmic internet, is still adequate in an era of amplification and virality.

This section examines New Zealand's Safe Harbour provisions, with a particular focus on the recommendations introduced by the FVPC Amendment Bill (September 2021). The primary objective is to assess whether Safe Harbour provisions should be strengthened by introducing a legislative duty of care for online content hosts, considering both theoretical frameworks and practical implications. In particular, Lessig's modalities remind us that Safe Harbour regimes are not only about law but also about how platform architecture and market incentives structure risk, while Mill's harm principle and Waldron's dignity-based account raise questions about

[654] Caio C.V. Machado and Thais Helena Aguiar "Emerging Regulations on Content Moderation and Misinformation Policies of Online Media Platforms: Accommodating the Duty of Care into Intermediary Liability Models." (2023) Business and Human Rights Journal 8 at 246-247.

whether immunity should persist when platforms facilitate foreseeable harms. Alongside these theoretical perspectives, this section also offers a comparative analysis of analogous provisions in other jurisdictions, providing a comprehensive understanding of Safe Harbour regulation and its limits.

Safe Harbour provisions apply to many kinds of internet intermediaries, including search engines, online auction sites, video-sharing services, and social media platforms.[655] They are a central part of how governments regulate online content. Because online content is so vast and constantly changing, deciding who should be legally responsible for harmful or illegal material created by third parties is not simple.[656] One useful way to think about this problem is through Barzilai-Nahon's[657] idea of "network gatekeeping". This theory shows that intermediaries do not only passively transmit information, but they also control and shape what flows through their networks.[658] If we accept this, then Safe Harbour protections should be questioned. Platforms that exercise such gatekeeping powers look less like neutral conduits and more like publishers, which suggests they may need stronger responsibilities than those currently imposed.

The concept of Safe Harbour is not unique to internet law. It also appears in company law, where it protects directors from personal liability if they meet certain conditions. For example, during the Covid-19 pandemic,[659] Safe Harbour legislation was introduced to shield directors from liability for breaches of their general duties.[660] This safeguard worked on the basis that

[655] Dan Svantesson *Internet Jurisdiction Global Status Report 2019: Key Findings* (2019) at 57.
[656] Laura C. Rodgriguez Rengifo "Liability of Internet Intermediaries: Participative Networking Platforms and Harmful Content" (LLM University of Wellington, 2016), at 4.
[657] Karine Barzilai-Nahon "Toward a theory of network gatekeeping: A framework for exploring information control" 2008 59 Journal of the American Society for Information Science and Technology 1493 at 1494.
[658] Rengifo, above n 656, at 5.
[659]Companies Act 1993, s 138B. This section has since been repealed on the close of 31 May 2022.
[660] Fiona MacKinnon "Safe Harbour from Insolvency Duties To Expire on 30 September 2020" (2020) <https://kindrik.co.nz/blogs/safe-harbour-from-insolvency-duties-to-expire-on-30-september-2020/>.

the Companies Act would not treat those duties as breached, provided that the directors satisfied the criteria set out in the legislation.[661]

This shows that the idea of Safe Harbour can be transposed to other areas, including social media platforms, internet service providers (ISPs), and internet protocol address providers (IPAPs). These actors function as intermediaries, hosting environments where users supply and share content. If Safe Harbour protects company directors in carrying out their roles under specific conditions, it may also be justified as a way to regulate intermediaries who enable, but do not directly produce, online content. From Lessig's perspective, Safe Harbour reflects the interaction of law and market incentives: it shields innovation by reducing legal risk, but it may also weaken other modalities such as norms and "code" if platforms rely too heavily on immunity. Murray's concept of dynamic regulation further reminds us that such protections cannot remain static and must adapt to contexts where algorithms actively amplify harm.

5.8.2 Safe Harbour elsewhere

In the European context, a thematic approach provides valuable insights into Safe Harbour provisions, particularly those outlined in the EU legal framework. Articles 12-14 of Directive 2000/31/EC establish limited liability protections for service providers regarding third-party content.[662] As noted earlier, several Member States, including France and Germany, have taken steps to refine and strengthen these protections in response to the changing risks of digital communication.

[661] MacKinnon, above n 660.
[662] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') 2000 OJ L/178.

The EU's e-Commerce Directive follows a horizontal approach, meaning a single legislative framework applies across all intermediaries, regardless of the specific type of content they host.[663] In 2016, the European Commission and major IT companies, including Facebook, YouTube, Twitter, and Microsoft, announced a Code of Conduct.[664] This initiative shows how Safe Harbour protections are now being supplemented by voluntary commitments to curb the spread of online hate speech.[665] From the perspective of Lessig's modalities, the Code illustrates how law interacts with norms (public pressure) and corporate "code" (platform design) to produce regulatory outcomes. Murray's theory of dynamic regulation also helps explain this development: instead of relying solely on formal legislation, the EU has encouraged flexible cooperation between regulators and platforms, adapting regulation to the fast-changing digital environment.

The key commitments in the Code of Conduct are as follows:

IT companies have committed to stronger initiatives to counter online hate speech, including implementing clear and effective review processes for user notifications.[666] These processes are intended to ensure that companies promptly assess whether content should be removed. The standards are set out in each platform's Rules and Community Guidelines, which explicitly prohibit material promoting violence and hateful conduct. When an individual or legal entity breaches these rules, a notification is sent to the company for review, with a commitment that such review should be completed within 24 hours.[667] From Lessig's perspective, this mechanism illustrates how law and norms are supplemented by code: the technical design of

[663] Rengifo, above n 656, at 14.
[664] European Commission "European Commission and IT Companies announce Code of Conduct on Illegal Online Hate Speech" (2016) Press Release <https://ec.europa.eu/commission/presscorner/detail/en/IP_16_1937>.
[665] European Commission, above n 664.
[666] European Commission, above n 664.
[667] European Commission, above n 664.

reporting systems and the 24-hour deadline act as regulatory tools in themselves. Murray's dynamic regulation theory also helps explain the significance of the arrangement, since the Code represents an adaptive, cooperative framework in which regulators and platforms share responsibility for responding to harms in near real time.

By way of counter-narratives, IT companies are committed to providing end-users with education and broader information about prohibited content, as set out in their Community Guidelines. They also undertook to make reporting mechanisms accessible so that individuals could notify companies of online hate speech or other harmful material. In addition, IT companies agreed to establish more effective communication flows with Member State authorities. This included creating national contact points to handle notifications and removal requests, with the aim of improving coordination and supporting law enforcement agencies in recognising and addressing harmful content. [668] From a regulatory theory perspective, these measures illustrate Lessig's modality of "norms," since the framework depends not only on law but also on shared practices and expectations to influence behaviour.

The EU's Safe Harbour principles are reflected in Article 42, which provides that "the limitations it places on the liability of online intermediaries exist because of their passivity in the curation and dissemination, and hosting of indexation of content on their platforms."[669] This framing suggests that social media platforms operate primarily as passive intermediaries, with limited responsibility for content curation and no direct editorial oversight. However, Professor Peter Coe of the University of Reading argues that these provisions may now be

[668] European Commission, above n 664.
[669] Peter Coe "The Draft Online Safety Bill and the regulation of online harms and hate speech: have we opened Pandora's Box?" (paper presented to BACL Annual Webinar: The Regulation of Hate Speech Online and Its Enforcement in a Comparative Perspective, London, 31 August 2021 2021).

obsolete.[670] He notes that while the directive assumes platforms are passive hosts, this view no longer reflects the reality of how they function, particularly given their algorithmic role in curating and amplifying user content.[671] From Lessig's perspective, this illustrates how "code" has altered the regulatory landscape: when platforms design systems that actively shape information flows, the rationale for treating them as passive intermediaries becomes harder to sustain. According to Dr. Christina Angelopoulos, Professor of Intellectual Property Law at the University of Cambridge, the "notice and action" procedure under Article 14(3) of the Digital Services Act (DSA) must be understood within the broader context of hosting Safe Harbour protections. She finds the connection between the two frameworks somewhat problematic, noting that the underlying logic is not entirely convincing.[672]

The DSA, however, marks an important shift in how the responsibilities of online intermediaries are conceived. It moves beyond the traditional assumption of platform passivity, advocating for a more active role in moderating harmful or unlawful content. The Act also differentiates obligations based on scale: intermediary services face minimal requirements, while "very large platforms" are subject to more comprehensive duties and regulatory oversight.[673] From Murray's perspective, this reflects an example of dynamic regulation in practice, where legal expectations evolve to match the risk profiles of different actors within the digital ecosystem.

[670] Coe, above n 669.
[671] Coe, above n 669.
[672] EuroISPA is the globally the largest association of internet service providers. EuroISPA "Liability of Intermediaries EuroISPA Recap of Past Event: Liability of Intermediaries" (podcast, 29 September 2021) EuroISPA < https://www.euroispa.org/2021/10/recap-of-past-event-liability-of-intermediaries/>.
[673] Thomas de Weerd and Jurre Reus "News Update - The Digital Services Act And The Digital Markets Act" (2021) <https://www.houthoff.com/insights/News-Update/News-Update-The-Digital-Services-Act-and-the-Digital-Markets-Act>.

5.8.3 Safe Harbour in New Zealand

New Zealand adopts a vertical regulatory approach, in which different statutes address specific legal issues concerning online content hosts and platforms. In contrast to the EU's horizontal model, which applies a single framework across intermediaries, New Zealand relies on separate statutes, such as the Harmful Digital Communications Act 2015 (HDCA), to assign distinct responsibilities depending on the type of harm.[674] Under this model, intermediaries are legally responsible for their own content, while liability for illegal user-generated content generally shifts to the individual user.[675] The vertical approach is illustrated in sections 23–25 of the HDCA, which govern the liability of online content hosts, with section 24 specifically setting out the safe harbour provisions.

According to section 24 of the HDCA, online content hosts in New Zealand can avoid liability for harmful material if they follow the statutory procedures once a complaint is received.[676] In practice, this provision can grant hosts immunity from both civil and criminal responsibility, even where content may otherwise breach the Films, Videos, and Publications Classification Act.[677] The only limitation is that section 24 does not apply to material that falls directly within the scope of the Classification Act, which deals primarily with objectionable and offensive publications.

As a general rule, section 24 of the HDCA requires an online content host to notify the author of the disputed material within 48 hours of receiving a complaint.[678] If the author cannot be

[674] Rengifo, above n 656, at 15.
[675] Ministry of Justice "Safe harbour provisions | New Zealand Ministry of Justice" (2021) <https://www.justice.govt.nz/courts/civil/harmful-digital-communications/safe-harbour-provisions/>.
[676] David Harvey *internet.law.nz* (Fifth ed, LexisNexis NZ Limited, 2023), at 216.
[677] Harvey, above n 676, at 217.
[678] Harmful Digital Communications Act 2015, s 24(2) and see Chapter 2.2 of this thesis.

identified, the host must promptly remove or disable access to the content. Where the author is known, they may issue a counter-notice either consenting to removal or opposing it.[679] If the author contests removal, the host is obliged to retain the material and inform the complainant of the author's decision.[680] With the author's consent, the host may also release limited personal information to the complainant. If no counter-notice is filed within the timeframe, the host must take down the content.[681] This framework is intended to secure swift resolution while balancing the rights of complainants and content creators.

Recent events in Christchurch have prompted New Zealand to reassess its legal framework for online content regulation, particularly in response to evolving digital threats. As part of this initiative, the Government is looking at introducing further legislation to combat online hate speech. It is looking particularly at updating the Classifications Act as a part of broader efforts to combat online hate speech.[682]

The Christchurch terrorist planned his attack using social media as a tool to livestream and the shootings were circulating on social media for 17 minutes.[683] These events intensified public debate about whether New Zealand's regulatory measures are sufficient to address online hate speech. In response, the Government introduced the Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2020 to strengthen the Films, Videos, and Publications Classification Act

[679] Harmful Digital Communications Act 2015, s24(2)(a)(ii).
[680] Harmful Digital Communications Act 2015, s24(d).
[681] Harmful Digital Communications Act 2015, s24(e).
[682] Note: this reflects one of the key mandates of the previous Labour Government. This was written prior to the new Coalition Government taking office in 2023, and it's uncertain whether this objective aligns with the current government's strategies.
[683] The Classification Office - Te Mana Whakaatu "Christchurch Mosque Attack Livestream" (2019) <https://www.classificationoffice.govt.nz/news/featured-classification-decisions/christchurch-mosque-attack-livestream/>.

1993 (FVPC).[684] At the same time, the Royal Commission of Inquiry into the Terrorist Attack on Christchurch Mosques on 15 March 2019 was tasked with investigating the circumstances of the attack and presenting recommendations to Parliament.[685] As noted in Chapter 2, the FVPC Amendment Bill sought to extend censorship powers by enabling the Chief Censor to issue takedown notices and by paving the way for the development of a web-filter. These proposals, however, have been contested on free speech grounds due to concerns about the potential for abuse of power.[686] From a Mill-ian perspective, this debate illustrates the need to distinguish between offensive content and content that causes real harm, while Breyer's proportionality framework highlights the importance of ensuring that measures like web-filters are calibrated to address serious threats without unnecessarily restricting lawful expression.

In short, by virtue of the FVPC Amendment Bill, "Safe Harbour" provisions of section 24 Harmful Digital Communications Act 2015 ("HDCA") will not apply.[687] The Amendment Bill explicitly renders section 24 inapplicable, shifting the responsibility onto online content hosts to take reasonable steps to remove objectionable material from their platforms. This represents a move away from reliance on intermediary immunity towards a model that places affirmative obligations on hosts, raising important questions about whether such duties amount to a statutory duty of care.

The Select Committee adjusted the FVPC Amendment Bill in its Final Report, limiting the relevance of the HDCA's safe harbour provisions in sections 23-25 to the newly inserted Part

[684] Refer to Chapter 2.2.
[685] Royal Commission of Inquiry into The Terrorist Attack On Christchurch Mosques On 15 March 2019, above n 261.
[686] Tan, above n 524.
[687] Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2020

7A.[688] This new part (sections 119A–119K) establishes a framework for takedown[689] notices directed at objectionable online publications. Section 119A defines the terms used in Part 7A, while section 119B sets out its scope and regulatory application.[690] Sections 119C and 119D outline the procedure and required contents of takedown notices, and section 119E obliges online content hosts to comply with them. [691] Importantly, section 119F shields officials from liability when issuing notices, while section 119G protects hosts that comply. [692] The remaining provisions address enforcement (s 119H), remedies and costs (s 119I), review mechanisms (s 119J), and reporting obligations (s 119K)[693]. Collectively, these reforms shift the burden onto online content hosts to take reasonable steps to remove objectionable material, embedding this responsibility within a statutory process. From Lessig's perspective, this is an example of "law" reshaping the regulatory environment of online speech, while Breyer's proportionality principle highlights the importance of ensuring that takedown obligations are balanced so they do not unduly restrict legitimate expression.

This raises a critical question: Should New Zealand follow the UK and Australia in imposing a statutory duty of care on social media platforms? Or does its current Safe Harbour framework (supported by self-regulation) strike the right balance between platform accountability and freedom of expression?

---

[688] Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2020 (select committee report) at 8.
[689] Takedowns, see Chapter 3.3 Disabling Access to Illegal Online Content - A Strategy for Online Hate Speech Regulation
[690] Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Act 2021, Part 7A, s 119A and s119B.
[691] Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Act 2021, s119C, s119D and s119E.
[692] Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Act 2021, s119G
[693] Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Act 2021, s119H, s119I, s119J and s119K.

The comparison highlights a key tension in New Zealand's regulatory approach. On the one hand, Safe Harbour provisions encourage innovation and protect intermediaries from excessive liability. On the other, they risk shielding platforms from responsibility for foreseeable harms, particularly when algorithms amplify harmful content. From Mill's perspective, this tension reflects the challenge of distinguishing mere offence from genuine harm, while Waldron's theory reminds us that dignity and public assurance may be undermined if platforms are overly insulated. The central question, then, is whether New Zealand should continue to rely on Safe Harbour or move towards statutory duties of care, as in the UK and Australia.

These provisions illustrate both the scope of Safe Harbour protections in New Zealand (broad intermediary immunity where notice procedures are followed) and their limits (narrowed by exceptions for objectionable material and increasingly challenged as insufficient to address systemic harms). The following section addresses this question directly. Having outlined the scope and limits of Safe Harbour protections, Section 6.3 turns to comparative statutory duty of care frameworks in the United Kingdom and Australia.

5.9 Non-State and Hybrid Responses

This section evaluates non-state and hybrid mechanisms that sit alongside statutory tools: platform Community Guidelines, the EU Code of Conduct, the Global Internet Forum to Counter Terrorism, and the Christchurch Call. These arrangements deliver speed and shared expertise, but they raise questions about consistency, accountability, and over-removal. The analysis applies the Chapter 2 lenses to assess whether these mechanisms protect dignity and participation without placing excessive burdens on lawful speech.

To combat the proliferation of terrorist and extremist content online, leading technology companies such as Microsoft, Facebook, Twitter, and YouTube launched the Global Internet Forum to Counter Terrorism (GIFCT) on 26 June 2017.[694] This initiative represents a unified industry effort to coordinate responses to extremist material in digital spaces.[695] GIFCT's work focuses on three areas: knowledge-sharing, research, and technological solutions. Through these, it develops best practices to support smaller platforms, funds research on emerging threats, and employs tools such as hash-sharing databases and machine learning systems to detect and remove harmful content quickly.

Following the Christchurch terrorist attacks in 2019, the urgency of coordinated responses gained global attention. New Zealand Prime Minister Jacinda Ardern and French President Emmanuel Macron launched the Christchurch Call to Action, a voluntary framework encouraging cooperation between governments, industry, and civil society to reduce terrorist and violent extremist content online. [696] Unlike binding law, the Call represents a hybrid governance model where non-binding initiatives supplement formal international obligations[697]. For New Zealand, a small jurisdiction dependent on global platforms, participation in such frameworks is not optional but essential.

GIFCT has since expanded its hash-sharing database to include nearly 390,000 items of extremist content, while its research arm, the Global Network on Extremism and Technology (GNET), has produced extensive analysis on online harms. [698] Its working groups have also

[694] Global Internet Forum to Counter Terrorism "Governance" (2020) <https://gifct.org/governance/>.
[695] Global Internet Forum to Counter Terrorism "Governance" (2020) <https://gifct.org/governance/>.
[696] Jacinda Ardern "Significant progress made on eliminating terrorist content online" (press release, 24 September 2019).
[697] Global Internet Forum to Counter Terrorism, "May 2019 Christchurch Call to Action" (2019) <https://gifct.org/about/story/#may-2019---christchurch-call-to-action >
[698] Global Internet Forum to Counter Terrorism*, Annual and Transparency Report 2023* (New York, GIFCT, 2024) <https://gifct.org/wp-content/uploads/2024/04/GIFCT-Annual-Report-2023.pdf.> at 3

developed practical tools such as playbooks on incident response and AI moderation. [699] These outputs illustrate Lessig's insight that "architecture", code and technical standards, operates alongside law and norms to regulate behaviour.

However, critiques remain. As Keller notes, broad systemic duties of care imposed on platforms risk incentivising over-censorship, particularly when "harmful content" is vaguely or politically defined.[700] Applied to GIFCT, this means that voluntary industry standards and technical enforcement tools may exceed what is strictly necessary under international human rights law. Without independent oversight, such models risk undermining public accountability[701]. Breyer's emphasis on proportionality is instructive: measures must pursue legitimate goals without imposing excessive burdens on free expression.[702] At the same time, Waldron's theory reminds us that protecting dignity and social assurance requires meaningful responses to extremist content.[703] The legitimacy of initiatives like GIFCT and the Christchurch Call therefore depends on whether they strike the right balance between safeguarding dignity and ensuring proportional limits on expression.

5.10 Lessons for New Zealand

International human rights standards often guide New Zealand's policy development, providing benchmarks for domestic legislation and ensuring alignment with global best

[699] Global Internet Forum to Counter Terrorism, *Red Team Working Group: Executive Summary* (New York, GIFCT, 2023) <https://gifct.org/wp-content/uploads/2023/11/GIFCT-23WG-1023-ExecSummary-1.1.pdf.>
[700] Daphne Keller "Broad Consequences of a Systemic Duty of Care for Platforms" (1 June 2020) The Center for Internet and Society <https://cyberlaw.stanford.edu/blog/2020/06/broad-consequences-systemic-duty-care-platforms.>
[701] Global Internet Forum to Counter Terrorism, *Red Team Working Group: Executive Summary* (New York, GIFCT, 2023) <https://gifct.org/wp-content/uploads/2023/11/GIFCT-23WG-1023-ExecSummary-1.1.pdf.>
[702] Breyer, above n 122, at 11-12 and 33.
[703] Waldron, above n 3, at 115.

practice. In the context of hate speech regulation, these standards also serve as reference points for assessing whether restrictions on expression remain necessary and proportionate. Incorporating these principles into law signals New Zealand's commitment to the rule of law at both the domestic and international levels, while also enhancing coherence within its legal system.[704] This reflects Stephen Breyer's proportionality framework, which emphasises that legal restrictions must not only pursue a legitimate aim but also avoid imposing excessive burdens on expression.

New Zealand's participation in international treaties carries with it obligations to give effect to their provisions through domestic law where necessary.[705] The Cabinet Manual not only offers guidance but also serves as a fundamental source of information concerning New Zealand's constitutional arrangements and treaty obligations. It records New Zealand's adherence to international agreements sauch as the ICCPR, underscoring the state's commitment to human rights and treaty compliance.[706] Through the Cabinet Manual, the executive branch recognises the importance of upholding international human rights commitments within domestic governance.[707] This framework is particularly significant in debates on hate speech, as it requires New Zealand to weigh international obligations under Article 19 and Article 20 of the ICCPR against its domestic constitutional traditions, a process that often calls for proportionality analysis.

Another key aspect of the legislative process in New Zealand is the scrutiny of bills against the New Zealand Bill of Rights Act 1990 (NZBORA). This Act is a cornerstone of New Zealand's

[704] Melhuish and Pacheco, above n 16.
[705] Law Commission, above n 403, at 116.
[706] Department of the Prime Minister and Cabinet- Te Tari O Te Pirimia Me Te Komiti Matua *Cabinet Manual* (2017) .
[707] Department of the Prime Minister and Cabinet- Te Tari O Te Pirimia Me Te Komiti Matua, above n 703.

legal system, ensuring that proposed legislation is consistent with fundamental human rights and freedoms. While the NZBORA is a domestic statute, its provisions often reflect international human rights standards, such as those found in the International Covenant on Civil and Political Rights (ICCPR). This alignment demonstrates the influence of international law on domestic rights protection, through formal application still depends on legislative intent and interpretation.[708] For instance, the New Zealand Law Commission has noted that international obligations impact domestic law-making processes across various domains, including human rights and trade agreements.[709] In practice, courts and policymakers often apply a proportionality framework when assessing restrictions on expression under NZBORA, echoing the ICCPR's requirement that limitations be necessary and proportionate. This doctrinal overlap highlights the tension between safeguarding free expression (as Mill would emphasise) and protecting dignity and social assurance (as Waldron argues), which sits at the heart of New Zealand's hate speech debates.

This interaction between international obligations and domestic law underscores the hybrid nature of New Zealand's approach: while treaty provisions are not directly enforceable without incorporation, they shape the interpretation of rights and the design of legislation. Against this backdrop, New Zealand's hate speech framework can be understood as both a product of domestic constitutional values and an evolving response to international human rights standards. The result is a framework that aspires to balance liberty and dignity through proportionality analysis but often struggles with under-enforcement. In theoretical terms, this reflects Breyer's caution that restrictions must be properly tailored, Mill's warning against silencing debate, and Waldron's insistence that dignity and social assurance provide a

[708] Law Commission "Introduction: A Globalising World" <https://www.nzlii.org/nz/other/nzlc/report/R34/R34-Introduc.html >
[709] Law Commission, above n 708, at 4.

legitimate ground for intervention. For New Zealand, this hybridity offers both flexibility and fragility: flexibility in adapting to global norms, and fragility in leaving gaps where domestic law falls short of international expectations.

5.11 Conclusion

This chapter has examined the regulatory frameworks employed by the European Union to manage illegal and harmful digital content, particularly focusing on the interplay between horizontal and vertical legislation.[710]

The e-Commerce Directive, as a horizontal regulation, provides a comprehensive framework that applies universally to digital platforms and content types.[711]Articles 12–14 of Directive 2000/31/EC, which outline limited liability provisions for platforms hosting third-party content, remain foundational, but the evolving digital landscape has exposed their limits.[712] The implementation of this effort demonstrates a conscious recognition of the ever-changing digital landscape and the increasing necessity for increased content filtering. However, it also raises questions about the adequacy of reactive systems in addressing rapidly disseminated harmful content.

Germany's NetzDG offers an instructive case study, reflecting both the promise and perils of content regulation. Unlike some regulatory proposals, NetzDG does not mandate proactive searches for illegal content, placing the onus on platforms to act upon being notified of

[710] Micova and de Streel, above n 469.
[711] European Commission "Combined Evaluation Roadmap/Inception Impact Assessment Digital Services Act package: deepening the Internal Market and clarifying responsibilities for digital services" (June 2020) < https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12417-Digital-Services-Act-deepening-the-internal-market-and-clarifying-responsibilities-for-digital-services_en>
[712] Wolfgang Schulz "Regulating Intermediaries to Protect Privacy Online – The Case of the German NetzDG" 2018

violations.[713] While this approach avoids the dangers of general monitoring, critics argue that the reliance on notification mechanisms incentivises over-removal and risks chilling lawful speech. The controversy underscores Breyer's warning that proportionality must guide regulation: restrictions must be effective in preventing harm without imposing excessive burdens on expression.

The regulation of online hate speech is at a pivotal juncture, where competing priorities-ensuring safety, preserving democratic freedoms, and managing global digital platforms-must be reconciled. Scholars like Wolfgang Schulz have highlighted the limitations of existing legal mechanisms in addressing online harms, underscoring the need for dynamic and adaptable solutions. [714] These critiques echo Mill's concern that state power may silence legitimate debate, and Waldron's insistence that dignity and assurance provide a necessary justification for intervention. Together, they illustrate the core dilemma of this chapter: preventing serious harm without eroding the democratic value of open discourse.

These comparative insights are not presented merely as descriptive surveys, but as lessons that directly inform the reform options or pathways for New Zealand. In particular, the EU's Digital Services Act demonstrates how statutory duties can embed accountability into platform governance, while the UK's Online Safety Act illustrates both the opportunities and risks of centralised regulatory oversight. For New Zealand, these examples highlight the importance of embedding proportionality into any reform: regulation must be effective in curbing harmful speech while preserving legitimate debate. Together, these examples underpin the

[713] Bundesministerium der Justiz und für Verbraucherschutz, above n 455.
[714] Schulz, above n 712.

recommendations in Chapter 6, where platform accountability, technological solutions, and legal frameworks are evaluated as pathways for reform in the New Zealand context.

Flowing on from the interaction between international obligations and domestic implementation, it is necessary to examine how New Zealand has developed its own legal framework to address hate speech. While the New Zealand Bill of Rights Act 1990 affirms the right to freedom of expression under section 14, this right is not absolute and is subject to "reasonable limits" prescribed by law under section 5. The balancing of expressive freedom against protections from harm is reflected in statutory provisions such as section 61 of the Human Rights Act 1993 and section 131 (incitement to racial disharmony), as well as in related provisions of the Crimes Act 1961.

Together, these statutes form the domestic foundation for regulating hate speech, and their interpretation demonstrates how New Zealand courts navigate the tension between free expression, human dignity, and social cohesion. The comparative insights developed in this chapter are significant because they show how proportionality, dignity, and accountability can be embedded into regulation. These principles directly inform the reform pathways discussed later in the thesis, ensuring that any New Zealand model is grounded in both international best practice and domestic constitutional values.

Chapter 6: Recommendations for Law Reform - A Statutory Duty of Care and Related Measures

6.1 Introduction

This chapter evaluates whether New Zealand should impose a statutory duty of care on social media platforms or continue to rely on self-regulation and post-hoc remedies. It draws on developments in the United Kingdom and Australia to assess how a statutory model could operate here, with a focus on mitigating online hate speech while safeguarding freedom of expression.

As outlined in Chapter 2, effective regulation cannot be understood in purely legal terms. Behavioural theories (e.g., the Online Disinhibition Effect; Social Learning Theory) show how platform design lowers social restraints and rewards hostility. Lessig's four modalities explain how law, markets, norms, and code jointly regulate speech. These insights ground the analysis that follows and justify a focus on platforms as both private governors and technological architects.

Normative theories of speech further frame the inquiry. Mill's harm principle provides a threshold for when state intervention is justified, while Breyer's proportionality analysis emphasises the need for regulation that is carefully balanced against free expression rights. Meiklejohn's civic republican perspective stresses that speech regulation must preserve democratic participation, and Waldron's dignity-based approach explains why unchecked online hate corrodes public assurance and equality. Considered as a whole, they supply the evaluative lens for the chapter's inquiry into whether New Zealand should impose a statutory

duty of care on platforms to mitigate hate speech while preserving the conditions of free and democratic discourse.

In addressing the overarching research question*, How effective is New Zealand's legal framework in addressing hate speech and harmful content on social media?,* this chapter sets out the principal law-reform recommendation to the thesis. It argues that a statutory duty of care and enhanced platform accountability offer the most effective means of remedying the deficiencies in the current approach. By analysing these reforms, the chapter contributes to developing a coherent and proportionate framework for combating online hate speech while safeguarding democratic freedoms.

Social media platforms such as Facebook and Twitter/X have become central spaces for civic discourse. Yet the same design features that facilitate connection also amplify harmful speech. The behavioural dynamics identified earlier, particularly the influence of anonymity and reward-based algorithms, help explain why social restraint weakens online. These conditions demonstrate that platform accountability cannot rest on private moderation alone and instead requires structured legal oversight, such as a statutory duty of care.

The question of accountability for online hate speech is complex. Responsibility may lie with individual users who generate harmful content, with platforms that design and moderate online spaces, or with the state through criminal law and regulatory enforcement. Determining the appropriate balance among these actors remains central to building an effective and proportionate response. Social media companies have introduced a range of measures to address harmful content, including community guidelines, reporting tools, and algorithmic detection systems. However, as discussed in Chapter 5, comparative experience in other

jurisdictions shows that these voluntary or self-regulated approaches remain inconsistent and unevenly enforced. This reinforces the need for clearer legal standards and external oversight in New Zealand.

Theoretical work helps explain why this problem is not easy to solve. Lessig has argued that regulation online does not come only from law but also from social norms, markets, and the "code" of the platform itself. In other words, the design of the software is already a form of regulation.[715] Murray develops this further by suggesting that regulation works best when it fits with the way digital systems already operate.[716] These perspectives show why platform design and accountability must be part of the legal discussion, not an afterthought.

Other theories also guide how limits on speech can be justified. Mill's harm principle sets a threshold: restrictions should only apply when real harm, not mere offence, is at stake.[717] Breyer speaks of proportionality, reminding us that restrictions must be carefully balanced with freedom of expression.[718] Meiklejohn sees speech in democratic terms, stressing that all voices should be able to participate in public debate. Waldron focuses on dignity, pointing out that when hate speech is unchecked, people lose the assurance that they belong in society. Taken together, these ideas give this chapter a framework to examine whether New Zealand law should impose clearer duties on social media platforms.

The chapter proceeds as follows. Section 6.2 introduces the statutory duty of care, distinguishing it from common-law negligence and explaining its application in the digital context. Section 6.3 examines how duty-of-care models are developing in the United Kingdom

[715] Lessig, above n 84, at 664.
[716] Murray, above n 134, at 240.
[717] Mill, above n 43, at 80.
[718] Breyer, above n 122, at 22.

and Australia. Section 6.4 then considers whether New Zealand should adopt a statutory duty of care, assessing how such a model would interact with existing frameworks such as the Harmful Digital Communications Act 2015. Section 6.5 concludes with the chapter's recommendations.

The legal classification of social media platforms, as intermediaries, services, or products, is decisive in determining whether they are treated as passive conduits with limited liability or as active publishers with heightened responsibility. Building on the analysis of New Zealand's Safe Harbour provisions in Chapter 5, this chapter evaluates how those rules balance platform immunity with user protection and argues that introducing a statutory duty of care would provide a clearer and more accountable framework. The discussion also examines whether New Zealand should adopt similar regulations to other jurisdictions, particularly given that Safe Harbour rules were drafted for an earlier internet era and may not address risks created by algorithmic amplification.

6.2 Duty of Care in the social media sphere

6.2.1 The Case for Reform

New Zealand should introduce a statutory duty of care requiring social media platforms to take reasonable and proportionate steps to prevent foreseeable harm caused by online hate speech and harmful content. The concept of duty of care, long recognised in negligence law, establishes responsibility for preventing foreseeable harm. While well established in physical and consumer protection contexts, it has yet to be fully applied in digital environments where algorithmic amplification and user-generated content create new forms of harm. This section outlines the conceptual foundation for a statutory duty of care, which is developed further

through the comparative analysis in Section 6.3. Together, these sections form the core of this chapter's argument for law reform in New Zealand.

It is important to distinguish between the common-law duty of care, developed through judicial decisions in tort law with the statutory duty of care imposed through legislation. The common-law duty traditionally applies to physical harm or economic loss but has not yet evolved to cover platform liability for online harms.[719] In contrast, statutory duties of care have been introduced in jurisdictions such as the United Kingdom[720] and Australia[721], where social media companies are required to take proactive measures to mitigate digital harms.

Social media platforms have historically been treated as neutral intermediaries, but the growing harm linked to online hate speech and algorithmic content promotion shows that this model is no longer adequate. New Zealand should follow international reforms by adopting a statutory duty of care that imposes proactive obligations on platforms to identify, assess, and mitigate foreseeable risks. This approach would provide a clearer and more consistent framework than reliance on evolving common-law principles, ensuring that regulation keeps pace with technological change and aligns with the New Zealand Bill of Rights Act 1990.

6.2.2 Theoretical and Normative Foundations

As discussed in Chapter 2, Lessig's concept of "code" and Waldron's dignity-based framework provide a useful foundation for assessing platform responsibility.[722] When algorithms curate and promote content, platforms act less as neutral conduits and more as active participants in

[719] Stephen Todd (ed) The Law of Torts in New Zealand (3rd ed, Wolters Kluwer Law International, 2020) at 37.
[720] *Online Safety Act* 2023 (UK).
[721] *Online Safety Act* 2021 (Cth)
[722] Lessig, above n 84, at 664 and Waldron, above n 3, at 115.

shaping online discourse. This reinforces the need to move beyond broad intermediary immunity toward a calibrated statutory duty of care that differentiates between passive hosting and active content promotion. Legal liability and due diligence should therefore operate together: due diligence ensures ongoing preventive action, while statutory liability provides accountability when those duties are ignored. This integrated approach would align New Zealand's legal framework with its human-rights commitments and with the proportionality principle underpinning the New Zealand Bill of Rights Act 1990.

This chapter also draws on Sandra Fredman's model of substantive equality to strengthen the normative foundation for the statutory duty of care proposed in this thesis. Fredman identifies four interrelated dimensions of substantive equality: redressing disadvantage, enhancing voice and participation, challenging stereotypes, and promoting institutional change.[723] Applied to the regulation of online hate speech and harmful content, these dimensions provide a principled framework for moving beyond reactive, content-based regulation towards structural reform. A statutory duty of care grounded in this approach would not only protect users from harm but also help build digital environments that enable full and equal participation, particularly for marginalised communities.

The Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression supports this position.[724] The Special Rapporteur emphasises that companies should assess the human rights implications of their products, take proactive measures to prevent harm, and establish clear due diligence procedures to ensure

[723] Fredman, above n 172, at 727.
[724] United Nations General Assembly Report of the Special Rapporteur on the promotion and protection of the freedom of opinion and expression UN Doc A/74/486 (9 October 2019) at [56].

accountability.[725] These expectations reflect the essence of the duty of care doctrine, which requires companies to take reasonable steps to prevent harm to those affected by their operations.[726] Waldron's dignity-based theory further underscores why such obligations matter: unchecked online hate speech erodes public assurance and equality. From a regulatory perspective, Lessig's framework highlights that law and "code" must work together to ensure platforms design systems that prevent harm rather than simply respond after it occurs.

6.2.3 Intermediary vs Active Platform

Integrating due-diligence obligations into corporate governance would enable social-media platforms to meet both ethical and legal expectations of responsibility. This convergence between legal accountability and corporate practice reflects a broader societal expectation that technology companies should act proactively to mitigate.

A widely accepted approach to platform liability is to classify social media platforms as intermediaries, meaning they function as conduits for interactions between users rather than as content creators or publishers.[727] Hutty, the Chair of EuroISPA's Intermediary Liability Committee, clarifies the principle of limited liability for intermediaries by distinguishing between: Services, such as telecommunications providers, which act as passive conduits for user communication; and Publishers, such as newspapers, which actively curate, edit, and disseminate content and therefore bear greater legal responsibility for what they publish.[728] This distinction supports a proportionate duty of care: the more actively a platform shapes or

[725] Report of the Special Rapporteur on the promotion and protection of the freedom of opinion and expression, above n 724, at [42].
[726] Report of the Special Rapporteur on the promotion and protection of the freedom of opinion and expression, above n 724, at [42].
[727] As discussed in Chapter 5, this intermediary model underpins New Zealand's Safe Harbour framework but increasingly fails to reflect the active role platforms play in shaping content exposure.
[728] EuroISPA, above n 672.

amplifies harmful speech, the stronger the justification for regulatory intervention. Lessig's modalities further show that platforms exercising control through design and algorithmic amplification are not passive conduits but active regulators of speech, confirming that intermediary immunity no longer reflects digital realities.

6.2.4 Comparative Illustration - EU and Algorithmic Transparency

European Commission President Ursula von der Leyen has emphasised the need for a unified legal framework to define the accountability of internet intermediaries. Speaking in the context of EU digital regulation, she argued: "…disseminate, promote, and remove content…(sic). We want the platforms to be transparent about how their algorithms work because we cannot accept that decisions that have a far-reaching impact on our democracy are taken by computer programs alone."[729] This highlights the European Union's position that social media platforms must assume greater responsibility by adopting a duty of care approach, recognising their dual role as both intermediaries (services) and active curators of content (products).[730]

Instead, a legal framework should define platform accountability for the distribution, advertisement, and removal of content. Central to this is algorithmic transparency: platforms must disclose how their systems prioritise and promote material, given that decisions with profound democratic consequences cannot be left solely to ambiguous computer programmes.[731] From Lessig's perspective, this underscores the regulatory force of "code,"

[729] Wishart, above n 629.
[730] Wishart, above n 629. The European Commission President also added that while there was a duty to disable Donald Trump's Twitter account, who was President of the United States of America at the time, following the events of 6 January 2022, it was at the same time the discretion to disable it should not have been entirely up to Twitter as it posed such an adverse effect on the freedom of expression.
[731] Wishart, above n 629.

while Breyer's proportionality analysis reminds us that such oversight must also respect legitimate expression.

As an additional point of clarification, the President of the European Commission stated that although it was necessary to disable Trump's Twitter account in response to the events that occurred on January 6, the decision to do so should not have been left entirely up to Twitter because it had such a negative impact on the right to freedom of expression.[732] From Meiklejohn's democratic perspective, leaving control of political speech to corporate actors undermines inclusive public participation.[733] At the same time, Waldron's focus on dignity underscores why unchecked hate speech cannot simply be tolerated.[734] The tension illustrates why a clearer statutory duty of care, combined with transparent oversight, may be preferable to reliance on discretionary platform governance.

6.2.5 Common-Law Analogy and Product Responsibility

The common law duty of care is developed through judicial decisions (as exemplified in *Donoghue v Stevenson*[735]) and establishes liability for foreseeable harm even without direct contractual relationships. Understanding this distinction is critical to evaluating how New Zealand should approach platform liability. However, the reach of these principles in digital contexts is uncertain. A statutory duty of care, modelled on the United Kingdom's Online Safety Act 2023, would offer clearer, enforceable obligations on platforms while maintaining consistency with New Zealand's rights framework.

[732] Wishart, above n 629.
[733] Meiklejohn, above n 104, at 26.
[734] Waldron, above n 3.
[735] *Donoghue v Stevenson* [1932] AC 562.

In *Donoghue v Stevenson*[736], the court held that manufacturers owe a duty of care to consumers to prevent foreseeable harm. Applied to digital environments, this reasoning suggests that platforms, like manufacturers, should ensure the environments they design do not foreseeably expose users to hate speech or other harm.

(i) Was there a legal obligation for Stevenson, as the maker of the ginger beer, to exercise care towards Donoghue, the consumer? - Similarly, should platforms be required to exercise care towards their users by ensuring algorithms and moderation systems do not expose them to foreseeable harm, such as online hate speech?

(ii) Was it pertinent that Donoghue did not personally buy the ginger beer and that her friend was the one who made the purchase? - In the platform setting, does it matter whether harm arises from content a user directly engages with or from harmful material shared by others in their network?

(iii) Did Donoghue have legal standing to pursue the suit against Stevenson? - By analogy, do social media users who suffer harm from online hate speech have sufficient legal standing and remedies against platforms under current law, or should a statutory duty of care provide clearer grounds for action?

By analogy, these questions guide how a statutory duty could clarify platforms' obligations toward users, regardless of the precise source of harm or contractual relationship. Platforms should therefore be treated as responsible for the digital environments they design rather than as neutral conduits. While the analogy to product liability is imperfect, it captures the normative expectation (consistent with Mill's harm principle and Waldron's concern for dignity) that platforms must exercise reasonable care to prevent foreseeable harm. Social media platforms should be regarded not merely as communication intermediaries but as hybrid products whose design and operation foreseeably influence user risk.

[736] *Donoghue v Stevenson* [1932] AC 562.

Accordingly, this chapter recommends that New Zealand recalibrate its current reliance on Safe Harbour provisions and move toward a statutory duty of care that reflects these realities. As Lord Atkin observed in *Donoghue v Stevenson*, a manufacturer owes a duty to take reasonable care for the safety of the ultimate consumer. Lord Thankerton later confirmed in *Bourhill v Young* that such relationships cannot be exhaustively catalogued, underscoring the doctrine's flexibility.[737] This adaptability supports extending the duty of care to digital platforms, which, although not tangible products, create environments that shape user interaction and risk. New Zealand should legislate a statutory duty of care requiring platforms to take proactive and proportionate steps to mitigate foreseeable harms, including the amplification of online hate speech.

Social media platforms complicate the product-service distinction. They function simultaneously as communication tools for users and as advertising services for businesses. Facebook and Instagram exemplify this dual role: while users engage socially, Meta monetises those interactions for commercial gain.[738] This hybrid character reinforces the need for a statutory duty of care, since platforms are active participants in shaping the online environment.

According to Daphne Keller[739], Director of the Program on Platform Regulation at Stanford's Cyber Policy Center, a statutory duty of care offers both opportunities and risks. One concern is that a universal standard risks becoming a one-size-fits-all approach, eliminating the nuanced

[737] *Donoghue v Stevenson* [1932] AC 562 per Lord Atkin. And *Bourhill v Young* [1943] AC 92 (HL)
[738] Joel Postman *SocialCorp : social media goes corporate* (New Riders, Berkeley, CA, 2009) at 55.
[739] Daphne Keller "The Center for Internet and Society" (2021) <https://cyberlaw.stanford.edu/about/people/daphne-keller>.

liability considerations that may be appropriate in specific cases.[740] Keller also notes that moving beyond reactive takedown procedures toward proactive prevention would represent a significant improvement in accountability.[741]

New Zealand should therefore strengthen its framework by introducing a statutory duty of care model, tailored to local conditions and consistent with the Bill of Rights Act 1990. Such reform would provide a transparent, proactive system that ensures platforms design their technologies with harm prevention in mind, rather than treating moderation as a secondary task. Building on this rationale, the following section 6.3 examines how statutory duties of care have been implemented in other jurisdictions, and what lessons they offer for New Zealand.

6.3 An exploration of the statutory duty of care approach in respective jurisdictions

Having established the theoretical and legal basis for a statutory duty of care, this section examines how similar frameworks have been implemented in other jurisdictions, particularly the United Kingdom and Australia, and what lessons these models offer for New Zealand. Social media platforms operate as internet intermediaries, yet their regulatory treatment differs widely across jurisdictions. Comparative experience provides useful insight into how statutory duties of care can enhance accountability and clarify platform responsibility. These frameworks also reveal deeper normative difference: the UK's approach reflects a rights-protective logic rooted in child safety and democratic participation, whereas Australia's model prioritises administrative efficiency and harm prevention within a utilitarian regulatory

[740] Daphne Keller "Broad Consequences of a Systemic Duty of Care for Platforms" (1 June 2020) The Center for Internet Society <https://cyberlaw.stanford.edu/blog/2020/06/broad-consequences-systemic-duty-care-platforms>.
[741] Keller, above n 740.

culture.[742] This contrast highlights that a "duty of care" can function as a moral concept in one jurisdiction and a managerial one in another, shaping how regulators balance free expression with public safety.

A key consideration in this analysis is the operation of New Zealand's Safe Harbour provisions, discussed in Chapter 5. These provisions shield intermediaries from liability for user-generated content under certain conditions. While they aim to balance innovation and protection, the growing harms linked to online hate speech highlight their limitations. As shown in Chapter 5, Safe Harbour rules have become increasingly difficult to justify when platforms use algorithmic systems that amplify harmful content. Their design choices now influence risk in ways comparable to producers of unsafe products. New Zealand's current approach therefore lacks a coherent mechanism for ensuring that immunity does not override user protection. A statutory duty of care would fill this regulatory gap by requiring platforms to identify and mitigate foreseeable harms while preserving lawful expression. This balance reflects the proportionality principle discussed earlier and grounds the argument for reform in practical regulatory terms rather than abstract theory.

In this context, social media platforms are not legally recognised as publishers in most jurisdictions, including New Zealand. However, their active role in curating and amplifying content places them closer to a category of "hybrid intermediaries." This functional reality blurs the traditional distinction between passive hosts and active speakers. Rather than reclassifying platforms as publishers, New Zealand law should strengthen the obligations that accompany their intermediary status. Under the current notice-and-takedown procedures,

[742] Sam Alexander, "A Uniquely Australian Approach: A Thematic Analysis of the Normative Foundations of Australia's Approach to the Regulation of the Internet" (2022) 43(1) Adelaide Law Review 345 at 356.

platforms are generally required to act only after receiving complaints, which limits responsiveness and places the burden on victims of harm. A statutory duty of care would supplement these procedures by requiring preventive risk assessment and transparent systems for detecting and addressing harmful content before harm occurs. Such a framework would not equate platforms with publishers but would impose proportionate responsibilities reflecting their control over digital environments.

From a regulatory perspective, Lessig's point that "code" and design choices are not neutral remains relevant: platform architecture already governs user behaviour, so regulation must ensure this power is exercised responsibly. Accordingly, New Zealand's framework should be clarified to ensure that Safe Harbour protections do not extend to situations where platforms algorithmically amplify harmful content. Statutory reform should make explicit that active content curation engages a corresponding duty of care.

Across major jurisdictions such as the United States, the European Union, Australia, and New Zealand, social media platforms are generally classified as intermediaries rather than publishers.[743] However, the degree of responsibility imposed on intermediaries varies considerably. The United States maintains broad immunity through Section 230 of the Communications Decency Act, whereas the European Union's Digital Services Act and Australia's Online Safety Act impose conditional and proactive duties of care. New Zealand's Safe Harbour provisions under the Harmful Digital Communications Act 2015 fall closer to the U.S. model, offering limited oversight and relying on notice-and-takedown mechanisms. This variation does not reflect confusion about classification but rather a policy choice about how far legal systems are willing to extend intermediary responsibility. From Murray's

[743] EuroISPA, above n 672.

perspective, this diversity underscores the challenge of "dynamic regulation," where effective governance requires coherence between legal frameworks, industry standards, and civil-society expectations rather than uniform labels.

Given that social media platforms are recognised as intermediaries rather than publishers, the key policy question is not how to classify them but how far their intermediary obligations should extend. As discussed in Section 6.2, common-law analogies have limited reach in the digital environment. What now matters is how statutory models operationalise a duty of care within the intermediary framework.

The following comparative analysis examines how two jurisdictions, the United Kingdom and Australia, have adopted distinct approaches to platform accountability. The United Kingdom has introduced an explicit statutory duty of care through the Online Safety Act 2023, while Australia has developed a functionally similar framework through the Online Safety Act 2021, enforced by the eSafety Commissioner. Each model provides useful insight into how proactive, risk-based regulation can enhance accountability and align platform responsibilities with the protection of users' rights, while maintaining respect for freedom of expression.

6.3.1 United Kingdom

The United Kingdom has introduced legal obligations on social media platforms through the Online Safety Act 2023, which imposes a statutory duty of care to reduce online harm. Originally proposed as the Online Harms Reduction Bill, the legislation underwent extensive parliamentary scrutiny before enactment.[744] The Act expands the regulatory powers of

[744] Lorna Woods "The duty of care in the Online Harms White Paper" 2019 11 Journal of Media Law at 6.

OFCOM, which is now responsible for ensuring compliance and imposing penalties for non-compliance.[745]

The Act adopts a risk-based approach, requiring platforms to assess and mitigate risks to users, particularly in relation to harmful content such as hate speech, cyberbullying, and illegal material.[746] This model is inspired by health and safety law, where regulated entities must take proactive steps to minimise foreseeable harm.[747] Social media platforms are therefore required to implement systems and processes that prioritise user safety, moving beyond reliance on reactive moderation policies.

To enforce these obligations, the Act expands OFCOM's role as an independent regulator with responsibility for developing codes of practice in consultation with industry stakeholders. These codes provide guidance on risk mitigation strategies, content moderation, and platform accountability, thereby embedding safety by design principles into digital spaces.[748] OFCOM's expanded mandate is significant because it demonstrates how a statutory regulator can translate abstract duties of care into enforceable standards.[749] It provides a coordinated framework where risk assessment, transparency, and accountability operate together under public oversight rather than corporate discretion. For New Zealand, this model illustrates how regulatory authority could be consolidated potentially through an empowered Netsafe or a purpose-built online safety commissioner; to ensure consistency and independence in platform governance.

---

[745] Secretary of State of Digital, Culture, Media and Sport, *'Consultation Outcome - Online Harms White Paper: Full government response to the consultation'.* <https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response.>

[746] William Perrin and Maeve Walsh Lorna Woods "Draft Online Harm Reduction Bill" (2019) <https://www.carnegieuktrust.org.uk/publications/draft-online-harm-bill/>.

[747] Perrin, above n 746.

[748] House of Lords, Select Committee on Communications, Parliament of United Kingdom, *Regulating in A Digital World*, 2nd Report, Session 2017-19.

[749] Department for Science, Innovation and Technology "Policy Paper - Final Statement of Strategic Priorities for Online Safety" (2025) < https://www.gov.uk/government/publications/statement-of-strategic-priorities-for-online-safety/final-statement-of-strategic-priorities-for-online-safety >

The Online Safety Act 2023 adopts a "safety by design" methodology that establishes statutory duties of care for online platforms, particularly in relation to illegal and harmful content. While earlier drafts of the legislation framed this more broadly under the Online Harms White Paper, the final Act now imposes specific, legally enforceable duties on user-to-user and search services to take "proportionate measures" to prevent individuals from encountering illegal content and to protect children from harm.[750] These statutory duties, though narrower than a general tort-based duty of care, reflect the same preventive logic: platforms must identify and mitigate foreseeable risks through risk assessments, transparency reports, and safety by design measures overseen by OFCOM.

This model parallels the proactive obligations imposed on employers under the Health and Safety at Work Act 1974, where responsibility lies in anticipating and minimising harm before it occurs.[751] For social media companies, this means embedding risk management and content-moderation systems into the design of their platforms to reduce the spread and amplification of hate speech and related harmful content. The Act therefore marks a clear shift from reactive takedown procedures to preventive governance. For New Zealand, the lesson is that a statutory framework need not replicate the UK's model wholesale but should incorporate its core principle: legally enforceable duties of care that require proportionate, forward-looking measures to protect users while preserving lawful expression.

A key strength of this regulatory model is its explicit recognition that hate speech constitutes harm and requires proactive intervention. Under sections 9 to 12 of the Online Safety Act 2023,

[750] *Online Safety Act 2023* (UK), ss 7-10.
[751] House of Lords, Select Committee on Communications, Parliament of United Kingdom, above n 786.

online platforms must carry out regular risk assessments, identify how their design features and algorithms may amplify harmful content, and implement proportionate mitigation measures.[752] OFCOM's draft Illegal Content and Child Safety Codes of Practice further elaborate these obligations by requiring platforms to integrate harm-reduction mechanisms into their service design, moderation systems, and user-reporting processes.[753] Unlike the reactive notice-and-takedown procedures under New Zealand's Harmful Digital Communications Act 2015, the UK framework embeds a forward-looking duty that obliges platforms to prevent foreseeable harms before they occur. This statutory model therefore represents a shift from post-harm response to risk prevention. For New Zealand, adopting a comparable statutory duty of care would move the regulatory focus from user complaints toward platform accountability, ensuring that safety measures are built into system architecture rather than left to voluntary practice.

While the Online Safety Act 2023 represents a major step toward statutory regulation, its approach remains cautious and fragmented. The Act replaces the former system of voluntary self-regulation with binding legal duties for platforms to conduct risk assessments and implement harm-reduction measures under sections 9-12.[754] However, these duties apply only to specific categories of "illegal" and "harmful-to-children" content, leaving a gap in relation to broader harms such as hate speech between adults. Consequently, platforms continue to exercise wide discretion in defining and enforcing their own content standards.

[752] *Online Safety Act 2023* (UK), ss 9-12.
[753] Online Safety Act Illegal Content Codes of Practice 2024: Explanatory Memorandum (UK Government, 2024) < https://www.gov.uk/government/publications/online-safety-act-illegal-content-codes-of-practice-2024-explanatory-memorandum>
[754] *Online Safety Act 2023* (UK), ss 9-12.

The Online Safety Act draws conceptually on the model first proposed by William Perrin, Maeve Walsh, and Lorna Woods in the Draft Online Harm Reduction Bill (2019), which outlined a statutory duty of care for online platforms. [755] For New Zealand, this suggests that relying on voluntary measures under the HDCA risks repeating the same inconsistencies, unless statutory reform establishes clear and enforceable standards that reflect the global and algorithm-driven nature of platforms. While that proposal envisaged a general duty to prevent foreseeable harm across all online activities, the Online Safety Act limits its focus to child safety and criminal content.

The UK framework therefore illustrates both the progress and the limits of current regulatory reform: it institutionalises accountability through OFCOM but stops short of establishing a comprehensive, general duty of care across all forms of harmful content. For New Zealand, this experience suggests that any future statutory duty should extend beyond child safety and criminal material to include systemic obligations addressing the amplification of hate speech and other foreseeable harms.

Implementing a statutory duty of care for social media platforms would establish a clear legal framework with enforceable obligations comparable to those found in health and safety regulation. Under such a framework, platforms would be required to identify, assess, and mitigate foreseeable risks associated with the design and operation of their services, including algorithmic amplification of harmful content. This approach would mirror the United Kingdom's Online Safety Act 2023, which mandates risk assessments and compliance plans overseen by OFCOM, and Australia's Online Safety Act 2021, which imposes proactive

[755] William Perrin "Government online harms proposals reflect Carnegie UK Trust work" (2021) Government online harms proposals reflect Carnegie UK Trust work <https://www.linkedin.com/pulse/government-online-harms-proposals-reflect-carnegie-uk-william-perrin?trk=public_profile_article_view>.

obligations through industry codes administered by the eSafety Commissioner. Non-compliance could attract significant financial penalties or operational restrictions, ensuring that platforms take preventative rather than reactive measures.[756]

Adopting a similar model in New Zealand would require recalibrating existing Safe Harbour protections under the Harmful Digital Communications Act 2015, which currently shield intermediaries from liability once they remove or disable access to harmful content. Replacing this reactive notice-and-takedown mechanism with a proactive duty to prevent foreseeable harm would strengthen accountability but would also raise concerns about potential overreach.

On the positive side, such reform would align platform responsibility with the principles of foreseeability and proportionality, ensuring that those who design and profit from digital environments bear responsibility for their risks. On the negative side, increased monitoring obligations could impose compliance burdens on smaller providers and risk chilling lawful speech if risk thresholds are defined too broadly. A well-designed statutory duty of care should therefore complement, rather than replace, Safe Harbour protections by conditioning immunity on demonstrable preventive action. This would move New Zealand toward a balanced framework in which platforms retain limited liability for user content but are legally obliged to reduce foreseeable harm through transparent systems and design safeguards.

Despite its promise, the Online Safety Act 2023 has been criticised for setting an enforcement threshold so high that practical implementation may prove difficult. The Act's "duty of care" obligations apply only where platforms fail to comply with OFCOM's codes of practice or neglect to complete a risk assessment, meaning enforcement depends on regulatory oversight

[756] Keller, above n 739.

rather than direct liability for individual harms. In practice, OFCOM must demonstrate systemic non-compliance before penalties can be imposed, and prosecutions are therefore rare.[757] As Coe notes, this approach captures only the most egregious failures while overlooking diffuse or emerging forms of harm, such as algorithmic amplification of hate speech or coordinated harassment that does not cross the criminal threshold.[758] This structural limitation produces a double effect: it is under-inclusive because it excludes evolving harms, yet over-inclusive because a broadly framed duty risks regulating minor or context-specific speech.

For New Zealand, this experience highlights the need to calibrate a statutory duty of care more precisely. A proportionate model would define harm in terms of foreseeability and risk of significant impact, rather than criminality, and would assign liability only where platforms demonstrably fail to apply reasonable preventive measures. Enforcement could occur through an enhanced regulatory mandate for Netsafe, rather than creating a new body. Netsafe could be empowered to issue compliance notices, require periodic risk-assessment reports from platforms, and impose civil penalties for systemic non-compliance, similar to the powers held by OFCOM in the United Kingdom and the eSafety Commissioner in Australia. In practice, this could be achieved by amending the HDCA to give Netsafe statutory investigative and enforcement powers, supported by transparent reporting obligations. Strengthening Netsafe's statutory authority would preserve institutional continuity while ensuring clearer accountability.

[757] Coe, above n 669.
[758] Coe, above n 669.

Breyer's proportionality approach provides a useful framework for such calibration: interventions must be effective in addressing foreseeable harm while remaining narrowly tailored to avoid chilling legitimate expression. A statutory duty built around transparency, proportionality, and systemic accountability would therefore achieve the preventive aims of the UK model while avoiding its procedural rigidity.

A further challenge in implementing a statutory duty of care lies in how "harm" is defined and operationalised. In the United Kingdom, the Online Safety Act 2023 and its earlier White Paper on Online Harms deliberately avoided providing a single definition of hate speech or harmful content, leaving scope for OFCOM to interpret these concepts through guidance and codes of practice.[759] This regulatory flexibility, while practical, also risks inconsistent enforcement and legal uncertainty. The issue is not the definition of hate speech itself but how such definitions determine the scope of the duty of care i.e. what risks platforms must anticipate, and what conduct triggers liability.

If New Zealand were to adopt a similar model, it would need to specify the threshold of harm that engages statutory duties, ensuring that regulation captures serious and systemic abuse without overreaching into protected expression. Breyer's proportionality analysis offers a framework for this calibration: obligations should be no broader than necessary to prevent foreseeable harm. In turn, Waldron's dignity-based approach explains why those duties must remain substantive, protecting individuals from degrading online environments while preserving democratic participation.

[759] Coe, above n 669 and Damian Tambini "The differentiated duty of care: a response to the Online Harms White Paper" 2019 11 Journal of Media Law 359 at 33.

In sum, while the United Kingdom's Online Safety Act 2023 does not expressly use the term "statutory duty of care", it effectively establishes one through its binding obligations on platforms to identify, assess, and mitigate risks of harm to users. This represents a shift from voluntary self-regulation toward enforceable, risk-based accountability. However, the framework's main limitation lies in its open-ended definition of "harm.", which leaves significant discretion to OFCOM and risks inconsistent application across platforms. If "harm" is confined only to clear legal violations, then the framework may prove too narrow to address the amplification effects of hateful or degrading material that fall short of criminal thresholds. Conversely, an expansive definition could overreach and chill legitimate expression.

For New Zealand, this illustrates that any move toward a statutory duty of care must be accompanied by a clear articulation of scope and proportional enforcement mechanisms. Whereas the United Kingdom has opted for an explicit regulatory model grounded in proactive risk assessment, Australia has pursued a more fragmented yet increasingly interventionist approach, shaped by the Christchurch attacks and domestic debates on digital safety.

The UK's framework therefore represents a rights-oriented, institutionalised model of duty, while we will see in 6.3.2, Australia's experience, reveals state-led, harm-prevention paradigm grounded in administrative oversight rather than codified rules. The UK's model therefore illustrates both the promise and practical limits of a statutory approach grounded in regulatory oversight. Australia's experience, by contrast, demonstrates a more decentralised but increasingly interventionist response.

6.3.2 Australia

Australia has not adopted a formal statutory duty of care for social media platforms. However, it has introduced one of the most comprehensive statutory online safety regimes in the world through the Online Safety Act 2021, which imposes proactive obligations on digital platforms to reduce and remove harmful content.[760] While these duties are not expressed as a "duty of care," they perform a similar regulatory function by requiring platforms to take reasonable steps to prevent foreseeable harm, particularly to children and vulnerable users.[761]

The Act empowers the eSafety Commissioner to issue removal notices, enforce compliance, and oversee industry codes of practice.[762] This centralised enforcement model contrasts with New Zealand's more reactive notice-and-takedown approach under the Harmful Digital Communications Act 2015. Australia's framework therefore represents a move toward functional responsibility rather than legal reclassification, a model that could inform how New Zealand integrates proactive accountability without necessarily adopting the "duty of care" label.

This regulatory trajectory reflects what Alexander describes as a "uniquely Australian" approach to internet governance, grounded in pragmatic harm reduction, fairness, and democratic security rather than a rights-based or moralist framework.[763] These normative foundations explain why Australia's online-safety regime emphasises harm prevention and administrative oversight through the eSafety Commissioner, rather than broad declaratory

[760] *Online Safety Act* 2021 (Cth), ss 65, 77, 88, 109, 114, 119 and 128.
[761] At s 65.
[762] *Online Safety Act* 2021 (Cth), Subdivision C.
[763] Sam Alexander, "A Uniquely Australian Approach: A Thematic Analysis of the Normative Foundations of Australia's Approach to the Regulation of the Internet" (2022) 43(1) Adelaide Law Review 345 at 356.

duties of care. The model prioritises measurable intervention and procedural fairness over abstract free-speech balancing, aligning with the Australian tradition of egalitarian and utilitarian public regulation.

Yet, Australia's model remains primarily administrative as opposed to rights-based. Accountability is enforced through regulatory discretion, not judicially reviewable duties owed to individuals. This reflects a broader Australian regulatory tradition that emphasises on pragmatic governance and centralised oversight. As Alexander notes, Australian digital policy tends to rely on state-driven intervention justified by public interest harm reduction, rather than on participatory or constitutional rights reasoning.[764] The result is a framework that is efficient but vertically structured; with limited avenues for contesting the regulator's decisions (a contrast with the UK's more pluralistic model).

Since 2010, the Australian government has voiced concerns about the internet's role in disseminating harmful content, including pornography and violent material, particularly targeting young and vulnerable people.[765] In response, the Australian Joint Select Committee on Cyber-Safety was established in March 2010 to examine how young people could engage with digital technologies safely and with confidence, ethical awareness, and an understanding of potential risks. This early initiative marked the beginning of a regulatory trajectory that has since expanded into more robust statutory interventions.[766]

In response to the 2019 Christchurch terrorist attack, the Australian government introduced the Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019. This legislation

[764] Alexander, above n 763, at 366.
[765] Paula Pyburne and Rhonda Jolly *Australian Governments and dilemmas in filtering the Internet: juggling freedoms against potential for harm – Parliament of Australia* (8 August 2014) at 9.
[766] Pyburne and Jolly, above n 765, at 29.

imposes strict obligations on internet service providers (ISPs), social media platforms, and hosting services, requiring them to swiftly remove or report livestreamed or shared violent content.[767] Failure to comply would mean severe penalties, reinforcing Australia's commitment to preventing the spread of extreme and harmful material. The Act demonstrates how a direct legislative response to Christchurch was adopted abroad, providing a model of decisive state intervention where content poses a clear and immediate threat. For New Zealand, the Australian approach illustrates how targeted statutory powers, particularly the ability to mandate rapid removal and reporting, could strengthen the current reliance on post-harm remedies under the Harmful Digital Communications Act 2015.

Beyond this, the Australian government has broadened its regulatory scope to tackle various online risks, including child grooming, the regulation of online gambling promotions, and the criminalisation of non-consensual image sharing.[768] Additionally, it introduced the Australian Code of Practice on Disinformation and Misinformation (the Code), developed by the Digital Industry Group (DIGI), an independent industry body.[769] This voluntary framework encourages platforms such as Google, Facebook, Twitter/X, TikTok, and Microsoft to implement measures to combat online falsehoods while balancing free speech protections.[770] Having said that, voluntariness limits consistency; statutory baselines are needed for minimum performance.

[767] House of Representatives Select Committee on Social Media and Online Safety *Social Media and Online Safety* (March 2022) at 71.
[768] House of Representatives Select Committee on Social Media and Online Safety *Social Media and Online Safety* (March 2022) at 71.
[769] House of Representatives Select Committee on Social Media and Online Safety *Social Media and Online Safety* (March 2022) at 71.
[770] House of Representatives Select Committee on Social Media and Online Safety *Social Media and Online Safety* (March 2022) at 71. And Asha Barbaschow "Facebook, Google, Microsoft, TikTok, and Twitter adopt Aussie misinformation code | ZDNet" (2021) <https://www.zdnet.com/article/facebook-google-microsoft-tiktok-and-twitter-adopt-aussie-misinformation-code/>.

Unlike the European Union's *Digital Services* Act, which mandates compliance through legally binding transparency and due-diligence obligations, the Australian Code of Practice remains voluntary and depends on self-reporting. This limits consistency and accountability, as enforcement relies on industry goodwill rather than regulatory oversight.[771] The Code nevertheless provides an example of coordinated soft law in which platforms agree to shared principles of harm reduction and transparency.[772] For New Zealand, the contrast between the EU's binding duties and Australia's voluntary code underscores a key policy choice: whether to rely on cooperative frameworks or to legislate enforceable standards that compel compliance.

A key feature of the Code is its clear definition of harm, providing guidance on what constitutes detrimental content.[773] The Code distinguishes between misinformation, referring to the unintentional spread of false or misleading material, and disinformation, involving deliberate attempts to deceive or manipulate public perception. This distinction is important because it allows platforms and regulators to differentiate between negligence and intent, enabling more targeted responses to online falsehoods. By defining these categories, the Code establishes a structured framework for addressing harmful digital content while maintaining respect for freedom of expression. However, its voluntary nature limits consistency across signatories, since enforcement relies on self-reporting and peer review rather than statutory authority.

The Online Safety Act 2021, implemented alongside the Australian Code of Practice on Disinformation and Misinformation, establishes a comprehensive regulatory framework aimed

[771] Commonwealth of Australia Government Response and Implementation Roadmap for the Digital Platforms Inquiry "Regulating in the Digital Age" (2019) at 16.
[772] Commonwealth of Australia Government Response and Implementation Roadmap for the Digital Platforms Inquiry, above n 771, at 17.
[773] Commonwealth of Australia Government Response and Implementation Roadmap for the Digital Platforms Inquiry, above n 771, at 8.

at minimising online harm.[774] One of its most significant innovations is the creation of the eSafety Commissioner, an independent governmental regulatory authority tasked with overseeing compliance, enforcing penalties, and taking action against harmful content.[775] The Commissioner's centralised oversight marks a significant shift from earlier self-regulatory models, thereby ensuring that content moderation and harm-reduction obligations are subject to public accountability rather than industry discretion. This consolidation of regulatory authority has made Australia a leading example of a coordinated, state-driven approach to online safety governance.

A defining feature of the Online Safety Act 2021 is its proactive enforcement structure, which empowers the eSafety Commissioner to issue removal notices for harmful or unlawful content, including extreme violent material, cyberbullying, and child-exploitation content.[776] This marks a shift from reactive complaint-driven systems toward a model of regulated prevention.

Unlike the European Union's Digital Services Act 2022, which imposes horizontal obligations on platforms to conduct risk assessments and publish transparency reports, the Australian framework centralises these functions in a single regulator with direct enforcement powers. The focus is not on systemic reporting but on operational compliance; the Commissioner can order removal within specific timeframes and impose penalties for non-compliance. Although Australia does not frame these obligations as a "statutory duty of care," they perform a similar function: platforms are required to anticipate foreseeable risks, respond promptly to harmful content, and demonstrate that their internal systems are adequate to prevent recurrence. This

[774] eSafety Commissioner "Online Safety Act 2021 Fact sheet" (2021) <https://www.esafety.gov.au/sites/default/files/2021-07/Online%20Safety%20Act%20-%20Fact%20sheet.pdf>.
[775] eSafety Commissioner, above n 774.
[776] Katharine Gelber "A better way to regulate online hate speech: require social media companies to bear a duty of care to users" (14 July 2021) <https://theconversation.com/a-better-way-to-regulate-online-hate-speech-require-social-media-companies-to-bear-a-duty-of-care-to-users-163808>.

approach therefore operationalises the preventive logic underlying a duty of care through statutory powers rather than through a common-law framework. From Breyer's proportionality perspective, the model demonstrates how swift regulatory intervention can protect users from severe harm while still allowing for procedural safeguards that preserve lawful expression.

Under the Online Safety Act 2021, online platforms are required to develop and comply with industry codes of practice registered by the eSafety Commissioner. These codes impose legally enforceable obligations on platforms and service providers to detect, remove, and prevent illegal or harmful content, particularly relating to child exploitation, cyber abuse, and violent material.[777] Unlike the European Union's Digital Services Act 2022, which mandates self-assessment and transparency reporting, Australia's framework embeds regulatory co-design: industry drafts the codes, but the Commissioner must approve, register, and enforce them. This model combines flexibility with statutory accountability, ensuring that voluntary industry standards are backed by legal consequences for non-compliance.

From a regulatory perspective, this approach operationalises what Suzor describe as "responsive co-regulation," blending government oversight with industry expertise to manage online harms.[778] It also embodies the principle of proportionality emphasised by Breyer, imposing binding obligations that are targeted and proportionate to the level of risk, rather than imposing blanket censorship. The Australian experience therefore demonstrates that proactive online safety regulation can maintain flexibility while still ensuring enforceability, offering a pragmatic middle ground between self-regulation and full state control.

[777] eSafety Commissioner, above n 774.
[778] Nicolas Suzor and Rosalie Gillett Flew "Responsive co-regulation." In Terry Flew and Fiona R. Martin (ed) *Digital Platform Regulation* (Palgrave Macmillan, 2022) pp. 263–264 and eSafety Commissioner "Regulatory Guidance on Industry Codes and Standards" (2022) <https://www.esafety.gov.au/industry/codes>

Australia's approach to platform liability has evolved beyond content removal obligations to include legal accountability for user-generated content. A key development in this regard is the Social Media (Anti-Trolling) Bill 2022, proposed in response to growing concerns about anonymous defamation and online harm following the High Court's decision in *Fairfax Media Publications v Voller* [2021] HCA 27 (Voller).[779] The Bill would have reclassified social-media platforms as publishers rather than mere intermediaries, thereby removing their ability to claim the "innocent dissemination" defence for defamatory user posts.[780] It also required platforms to reveal the identity of anonymous users accused of defamation. Legal scholars and civil-society organisations criticised the proposal for targeting individual identity exposure rather than systemic platform duties, arguing that it misconceived the nature of online harm and risked undermining privacy and free expression.[781]

Although the Bill lapsed in April 2022 and was not reintroduced,[782] it remains instructive as a cautionary example of reactive, speech-restrictive regulation that contrasts with duty-of-care-based approaches. Rather than expanding liability through identity-unmasking powers, a New Zealand framework should prioritise risk-based obligations, transparency, and systemic accountability.

The Bill ultimately failed to pass Parliament.[783] The Senate Legal and Constitutional Affairs Legislation Committee found it was "not fit for purpose," warning that it might discourage

[779] Explanatory Memorandum House of Representatives "Social Media (Anti-Trolling) Bill 2022" (2022) <https://parlinfo.aph.gov.au/parlInfo/download/legislation/ems/r6831_ems_d8a044e1-2ac3-4f15-b90a-7cf5d57b4b2e/upload_pdf/JC004985.pdf;fileType=application%2Fpdf >.
[780] Explanatory Memorandum House of Representatives, above n 779.
[781] Alice Klein "Proposed anti-trolling law could be used to silence critics of the Australian government" *New Scientist* (12 November 2021) 252 (3364) 10.
[782] Senate Legal and Constitutional Affairs Legislation Committee, *Report on the Social Media (Anti-Trolling) Bill 2022* (April 2022).
[783] Parliament of Australia "Social Media (Anti-Trolling) Bill 2022" 2022) < https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bId=r6831>

defamation claims and misdirect resources away from genuine online-harm prevention.[784] Moreover, its reliance on court orders to unmask anonymous users raised both practical and privacy concerns.[785]

Despite its failure to pass, the Bill illustrates a wider shift in Australian regulation, from voluntary self-regulation, toward compelled intervention and greater legal exposure for intermediaries. It therefore underscores the spectrum of emerging global responses: while the United Kingdom embeds statutory duties through OFCOM oversight, Australia experiments with targeted, speech-focused mechanisms that reveal the limits of piecemeal reform. While the UK and Australia demonstrate two distinct pathways for enhancing platform accountability, both approaches reflect a broader shift towards a stronger platform responsibility. Yet the Australian experience shows that without a clear, statutory duty of care, regulation risks becoming fragmented and reactive.

In conclusion, the contrasting experiences of the United Kingdom and Australia reveal the spectrum of contemporary regulatory design. For New Zealand, these lessons reinforce the need for a proportionate, preventive framework that balances user protection with expressive freedom. Such a model should integrate systemic obligations, similar to the UK's risk-assessment approach, while avoiding the individualised, identity-based mechanisms rejected in Australia. Everything considered, the United Kingdom's and Australia demonstrate that a statutory duty of care is not a standard and uniform legal construct; instead, it is a contextually

[784] Parliament of Australia "Labor Senator' Minority Report" < https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Legal_and_Constitutional_Affairs/Anti-Trolling/Report/section?id=committees%2Freportsen%2F024900%2F79576>
[785] Business Standard "Australia's social media anti-trolling bill raises alarm for tech giants" (2022) <https://www.business-standard.com/article/international/australia-s-social-media-anti-trolling-bill-raises-alarm-for-tech-giants-122030700244_1.html >.

adaptive mechanism shaped by each jurisdiction's normative priorities; rights protection in the UK and administrative pragmatism in Australia.

6.4 Recommendation: Introduce a statutory duty of care for social media platforms in New Zealand

New Zealand should introduce a statutory duty of care for large social media platforms. Such a reform would move the current framework beyond reactive complaint-based regulation and establish clear, enforceable obligations to identify, assess, and reduce foreseeable risks of online hate speech and related harms. This model aligns with the preventive logic of the United Kingdom's Online Safety Act 2023 while remaining consistent with New Zealand's regulatory tradition of proportionate, flexible intervention.

The current framework, centred on the *Harmful Digital Communications Act 2015* and its Safe Harbour provisions, relies heavily on voluntary compliance.[786] While some content hosts, such as Radio New Zealand, act responsibly, this depends on institutional goodwill rather than enforceable obligation. Introducing a statutory duty of care would ensure that all major platforms meet minimum safety standards consistently, regardless of their size, resources, or internal policy.

On the other hand, Australia's Online Safety Act 2021 offers a different model of operationalising platform responsibility. Rather than introducing a single "duty of care," it imposes layered content-specific obligations and graduated enforcement powers administered

[786] Harmful Digital Communications Act 2015, s24.

by the eSafety Commissioner.[787] Platforms must remove class 1 (child sexual abuse and extreme violence) material within 24 hours of notice and respond to other harmful-content complaints under defined timeframes.[788] The Commissioner can issue removal notices, require transparency reports, and develop industry codes and standards tailored to service categories such as social-media, search, or hosting services.[789]

This framework contrasts with the UK's broader and more conceptual "duty of care." While the UK Act depends heavily on internal risk assessment and proportionality tests, Australia's system gives regulators direct enforcement levers and measurable timelines. However, it can be argued that it risks over-centralising discretion in the eSafety Commissioner and lacks an explicit balancing clause equivalent to the UK's freedom-of-expression safeguard .

For New Zealand, these comparative lessons point to the value of combining Australia's clear procedural duties and enforceable timelines with the UK's rights-based proportionality framework. A statutory duty of care could be supported by risk-assessment and transparency obligations but implemented through defined regulatory powers similar to those of the eSafety Commissioner.

6.4.1 Statutory duty of care for large social media services

This thesis draws an analogy between the common law duty of care and the proactive obligations recognised in human rights law. Both frameworks centre on the idea of foreseeable harm and positive responsibility. The comparison is not doctrinal but conceptual: it supports a

[787] *Online Safety Act* 2021 (Cth), s 28.
[788] *Online Safety Act* 2021 (Cth), 109-113A.
[789] Office of the eSafety Commissioner "Regulatory guidance" (2025) <https://www.esafety.gov.au/industry/regulatory-guidance#compliance-and-enforcement-policy>

model of platform regulation where entities exercising significant control over users' digital environments must take reasonable steps to prevent predictable harm to rights and dignity. The development of a statutory duty of care for social media platforms represents a turning point towards the case *Donoghue v Stevenson*, where the law first recognised that manufacturers owed duties to consumers beyond contractual relationships.[790] Just as Lord Atkin's neighbour principle marked a doctrinal shift in negligence, extending responsibility to foreseeable harms, a statutory duty of care for platforms would impose proactive obligations to address online hate speech. In practice, some hosts have already adopted self-regulatory measures. In the digital context, social media platforms occupy a similar relational position of control and foreseeability: their algorithmic design and moderation systems shape users' exposure to harmful content. The duty of care concept, therefore, provides a familiar and normatively grounded template for translating that relational responsibility into law.

For example, Radio New Zealand (RNZ)[791], proactively moderates comments on its Facebook page**.** While this shows how hosts can exercise responsibility, it is not legally mandated and depends entirely on institutional choice. While global platforms already operate moderation systems, RNZ's approach is significant because it demonstrates domestic application of co-regulation, where a New Zealand host applies both the HDCA and platform guidelines within its own editorial framework. While voluntary moderation by hosts (e.g., RNZ's comment practices) demonstrates responsible hosting, its scope remains uneven and unenforceable. A statutory duty would transform such isolated initiatives into a universal legal standard, ensuring that every major platform meets minimum safety obligations regardless of internal policy.

[790] *Donoghue v Stevenson* [1932] AC 562 per Lord Atkin.
[791] Radio New Zealand (RNZ) is an independent public multimedia organisation which is also a Crown entity pursuant to the Radio New Zealand Act 1995 (NZ). See https://www.rnz/about.

From Waldron's perspective, such obligations would protect dignity by ensuring that targeted groups retain public assurance of belonging. This view is echoed in the Royal Commission of Inquiry into the Christchurch Terrorist Attack, which emphasised the need for stronger legal safeguards against online hate.[792] Similarly, Lessig's modalities remind us that algorithmic "code" itself can entrench harm unless subject to oversight, as recognised in OFCOM's regulatory guidance under the UK Online Safety Act 2023.[793]

RNZ's moderation of its Facebook page demonstrates how self-regulation currently operates under the Harmful Digital Communications Act 2015.[794] By issuing visible warnings and disabling comments likely to attract abuse, and by applying Facebook's Community Guidelines, RNZ combines legal obligation with platform norms to manage harm.[795] Yet this model remains voluntary and inconsistent across hosts.[796] Unlike the United Kingdom's statutory duty of care framework discussed in section 6.3.1, which legally requires platforms to identify and mitigate online harms, New Zealand's system still depends on discretionary compliance. Adopting a similar statutory duty of care would transform voluntary moderation

[792] William Young and Jacqui Caine *Royal Commission of Inquiry into The Terrorist Attack on Christchurch Mosques On 15 March 2019* (November 2020), above n 287.
[793] Secretary of State of Digital, Culture, Media and Sport, *'Consultation Outcome - Online Harms White Paper: Full government response to the consultation'.* <https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response.>
[794] Harmful Digital Communications Act 2015
[795] Radio New Zealand "RNZ : Harmful Communications" (2021) <https://www.rnz.co.nz/harmful-communications>.
[796] As a platform host, RNZ encourages users to reflect on specific questions before publishing content: 'Ask yourself: would this offend someone? Is it defamatory? How would you react if someone else wrote the same thing?' This could be interpreted as RNZ's effort to fulfill its obligations under the HDCA or, alternatively, as a strategy to shift the 'duty of care' from the host to the user. Regardless, this example underscores the limitations of self-regulation in countering online hate speech, emphasizing the necessity of considering enforceable statutory obligations for key players in the social media landscape. The boundary is drawn when users violate two fundamental aspects: (i) the community guidelines of respective social media platforms and (ii) the host's Charter, which outlines its operational principles and policies for handling illegal content and hate speech. In RNZ's case, its Charter articulates the primary goals of the public radio company, emphasizing a "sense of responsibility" towards community interests and a commitment to accommodate or encourage those interests whenever possible. However, the pivotal question remains: Is self-regulation adequate to mitigate the impact of online hate speech in the realm of social media?

into enforceable standards, ensuring that protection against online hate speech does not depend on institutional goodwill but on clear legal accountability.

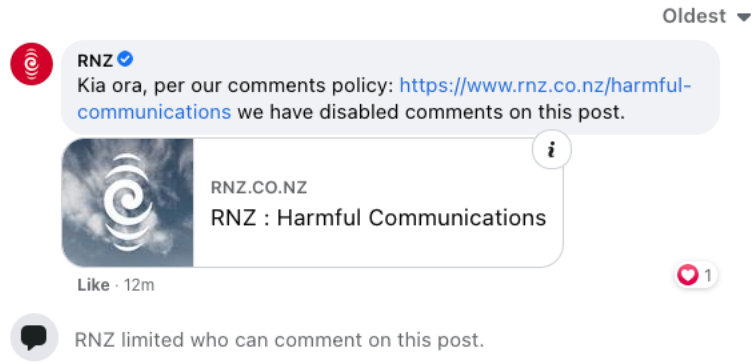

Figure 2: RNZ's Comment Section on Facebook

Social media platforms are best understood as internet intermediaries, entities that "bring together or facilitate transactions between third parties on the Internet," giving access to, hosting, transmitting, and indexing content or services originated by others.[797] This category includes Internet service providers, web-hosting companies, search engines, and participatory networking platforms.[798] While legally classified as service providers rather than products, their design and operational choices give them gatekeeping power over online expression. As UNESCO observes, intermediaries increasingly function as "chokepoints, arbiters or gatekeepers of expression," which means they are not neutral conduits but active participants in shaping digital communication.[799]

From a liability perspective, their classification as intermediaries rather than content producers has major implications: it determines whether they enjoy Safe Harbour protection or bear

[797] Karine Perset *The Economic and Social Role of Internet Intermediaries* OECD Digital Economy Papers No 171 (2010) at 9.
[798] Perset, above n 797, at 10.
[799] Rebecca MacKinnon et al *Fostering Freedom Online: The Role of Internet Intermediaries* (UNESCO, Paris, 2014) at 15-23.

proactive duties of care. As *Donoghue v Stevenson*[800] established, foreseeability of harm can justify the imposition of a duty; likewise, when platform's design choices foreseeably amplify hate speech, they should bear obligations to mitigate that risk. A statutory duty of care would codify this responsibility, ensuring that those who design and profit from digital environments take proportionate steps to prevent foreseeable harm. Imposing a statutory duty of care would not replicate tort liability but would establish enforceable regulatory standards for platform design, moderation, and transparency. It would shift the focus from individual fault to systemic prevention, aligning with the broader human rights duty to protect individuals from foreseeable harm caused by third parties.

This chapter therefore recommends legislating a statutory duty of care requiring platforms to take proactive steps to mitigate foreseeable harms, including the amplification of online hate speech. For New Zealand, this experience suggests that reforms should prioritise systemic accountability measures, such as risk-based duties and transparency obligations, over identity-based or punitive approaches. To ensure such a duty is effectively implemented and enforced, New Zealand also requires an independent regulatory mechanism with statutory powers to oversee compliance.

6.4.2 Independent regulation and enforcement

An independent online safety regulator should be established to oversee platform compliance with statutory duties of care. The Christchurch Call underscored the need for proactive, coordinated measures, urging social media companies to adopt stronger content moderation

[800] *Donoghue v Stevenson* [1932] AC 562.

policies and governments to introduce clear regulatory obligations. [801] In line with this, both the Helen Clark Foundation and the Royal Commission Inquiry into the Christchurch Terrorist Attack have proposed the creation of an independent body to monitor and enforce online safety standards. [802] The Royal Commission in particular called for stronger legislative protections to curb the spread of hate speech and extremist material online, while the Helen Clark Foundation emphasised the need for a specialist regulator with powers to impose penalties where platforms fail to prevent foreseeable harm.[803] Together, these reports establish a clear precedent for reform: New Zealand needs an independent, statutory regulator to move beyond voluntary and complaint-driven enforcement. Although initiatives such as Netsafe's voluntary Code of Practice have filled short-term gaps, they lack binding authority and remain dependent on industry goodwill. A dedicated statutory regulator would institutionalise accountability, ensuring that safety standards are applied consistently across platforms rather than left to voluntary compliance.

An independent regulator should be established to oversee compliance with statutory online safety duties and to enforce systemic accountability. The regulator would develop and approve industry codes of practice, conduct audits of high-risk platforms, and issue improvement or penalty notices for repeated non-compliance. It could also coordinate with existing bodies such as the Privacy Commissioner and the Classification Office to avoid duplication. Similar to Australia's eSafety Commissioner, this body should operate independently from ministerial control, with transparent reporting obligations and judicial review of its decisions. In New

[801] Claire Mason and Kathy Errington, 'Claire Mason and Katherine Errington "Anti-social media: reducing the spread of harm content on social media networks" (2021) https://helenclark.foundation/publications-and-media/anti-social-media/

[802] William Young and Jacqui Caine *Royal Commission of Inquiry into The Terrorist Attack on Christchurch Mosques On 15 March 2019* (November 2020), above n 287.

[803] Royal Commission of Inquiry Into The Terrorist Attack On Christchurch Mosques On 15 March 2019 "Part 3 - Hate speech – sections 61 and 131 of Human Rights Act 1993" (2021) < https://christchurchattack.royalcommission.nz/publications/comp/hate-speech-sections-61-and-131-of-human-rights-act-1993> at 51.

Zealand, this could take the form of an "Online Safety Commission," housed within but functionally independent from the Department of Internal Affairs, with statutory powers to monitor compliance, request transparency reports, and impose administrative penalties. Centralising oversight in this way would ensure that platform safety standards are applied consistently rather than left to voluntary goodwill.

The feasibility of such a model in New Zealand depends on striking the right balance between enforceability and proportionality. This can be achieved through a tiered regulatory design: large or high-risk platforms (for example, Meta, Twitter/X, TikTok, and YouTube) would be subject to full statutory duties, while smaller or niche services would be governed by lighter-touch, code-based obligations. A statutory duty of care should be expressly defined as a "reasonable steps" obligation, requiring risk assessment, mitigation, and transparency and not general monitoring or censorship, to preserve consistency with the New Zealand Bill of Rights Act 1990.

This approach would align New Zealand with the United Kingdom's preventive model while reflecting the proportional safeguards emphasised by Breyer's proportionality test and Lessig's responsive architecture. It would also avoid the pitfalls of Australia's over-expansive enforcement by embedding clear statutory limits and judicial oversight. In short, New Zealand should adopt a tiered, proportionate regulator with defined powers to enforce a statutory duty of care; firm in authority but flexible in scope. Effective regulation also requires clear procedural duties (risk assessment, transparency, and user redress) to ensure that the statutory framework operates in practice and maintains public trust.

6.4.3 Risk assessment, transparency, and user redress

The UK model's high-level risk taxonomy gives structure to the regulation of online harms but, in the absence of clear enforcement standards, leaves significant discretion to platforms and produces uneven compliance. The Online Safety Act 2023 (UK) requires major services to identify, assess, and mitigate online harms through a structured classification process.[804] However, while the framework defines categories such as hate speech, disinformation, and cyberbullying, it lacks prescriptive guidance on how risk assessments must be conducted or how "reasonable and proportionate" measures are to be judged. This indeterminacy has been one of the most frequently cited criticisms of the Act.[805] For New Zealand, this highlights a central design lesson: if a statutory duty of care is adopted, its enforcement standards must be more explicit to avoid replicating these weaknesses.

A statutory duty of care in New Zealand should therefore be operationalised through three interconnected obligations, risk assessment, transparency, and user redress, each designed to ensure that platforms act proactively and accountably while preserving freedom of expression under section 14 of the New Zealand Bill of Rights Act 1990.

6.4.3.1 Risk assessment

Platforms should be required to undertake annual and event-triggered risk assessments that identify foreseeable harms arising from system design, algorithms, and moderation processes. These assessments must include:

[804] *Online Safety Act 2023* (UK) ss 10-12; and Ofcom Guidance (2024).
[805] House of Lords Communications & Digital Committee "Free for All? Freedom of Expression in the Digital Age" (2021) <https://lordslibrary.parliament.uk/freedom-of-expression-online-communications-and-digital-committee-report/> .

- an evaluation of how platform architecture (such as recommender systems or virality tools) may amplify hateful or harmful content;
- identification of vulnerable user groups disproportionately affected by such harms; and
- the mitigation measures in place to prevent recurrence.

Risk assessments would be submitted to an independent online safety regulator (as proposed in section 6.4.2) and subject to audit for Tier 1 and Tier 2 services. This approach ensures accountability through documented foresight rather than post-harm enforcement.

6.4.3.2 Transparency

Platforms should also be bound by transparency duties to publish standardised metrics on their performance, including takedown response times, error-correction rates, reinstatement statistics, and volume of complaints handled. These metrics would be defined in secondary legislation or industry codes approved by the regulator. The aim is not to expose trade secrets but instead to allow public, academic, and regulatory scrutiny of how effectively platforms manage harm. Transparency also conditions Safe Harbour protection, which should remain available only to platforms demonstrating good-faith compliance with statutory benchmarks.[806]

6.4.3.3 User redress

Effective regulation depends on user trust. Platforms should therefore be required to implement accessible and fair redress systems. Users must receive clear reasons for moderation decisions,

[806] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') 2000 OJ L/178.

a defined period to appeal internally, and (where systemic failures occur) access to external review by the regulator. This model moves away from identity-unmasking or punitive approaches and focuses instead on system design and accountability, thereby preserving privacy while improving procedural fairness.

6.4.3.4 Enforcement and proportionality

The statutory framework should embed explicit enforcement standards to give practical meaning to the duty of care. These could include:

- fixed timeframes for risk reporting and appeals;
- audit obligations at set intervals for high-risk platforms;
- publication of compliance notices and penalty guidelines; and
- a categorised enforcement scale based on the platform's size and history of non-compliance.

Compliance could be assessed via measurable benchmarks similar to those applied under the UK Online Safety Act 2023. For example, platforms would be required to remove any manifestly illegal content withing a 24–48-hour timeframe, publish quarterly transparency reports that details takedown accuracy and user-appeal outcomes and resolve any internal appeals withing 14 days. Non-compliance that is persistent could attract escalating penalties; this ensures that statutory obligations translate into concrete and verifiable performance standards.

To balance enforceability with proportionality, New Zealand should adopt a tiered model. Very large or high-risk platforms (e.g., Meta, Twitter/X, TikTok, YouTube) would be subject to full statutory duties, while smaller or niche services would face lighter, code-based obligations.

This reflects Breyer's proportionality principle and avoids overreach, ensuring that intervention remains necessary and reasonable in a free-expression context.

The limits of New Zealand's current framework (its reactive tools, narrow protected characteristics, and reliance on platform goodwill) demonstrate that voluntary compliance is insufficient to address online hate. The *Harmful Digital Communications Act 2015* and the *Broadcasting Act 1989* offer some remedies but lack systemic oversight and clear risk-management duties. By contrast, the UK and Australian models show that statutory duties of care, combined with enforceable risk-assessment and transparency requirements, can strengthen user protection while holding platforms accountable.

In this light, New Zealand's duty of care should be designed as a preventive, evidence-based, and proportionate framework, supported by transparent reporting and accessible redress mechanisms. When viewed together, these components would transform online-safety regulation from a reactive complaint-driven model into a proactive system grounded in accountability, transparency, and public trust.

While self-regulation remains the primary mechanism for online safety in New Zealand, its limits are clear. Organisations such as Netsafe play an important role through education, digital-safety research, and public reporting of harmful content, and their development of the Aotearoa New Zealand Code of Practice for Online Safety and Harms represents a valuable industry-led step. However, without legally binding obligations or enforcement powers, the Code depends on platform goodwill and lacks the consistency required to ensure long-term protection. A statutory duty of care, implemented through the risk-assessment, transparency, and redress obligations outlined above, would give these voluntary initiatives legal force. This

hybrid model, combining co-regulation with enforceable standards, would therefore preserve the collaborative strengths of Netsafe's approach while embedding accountability across the sector.

6.4.4 Expanding protected characteristics in New Zealand hate-speech law

A complementary reform concerns the limited scope of protected characteristics under New Zealand's hate-speech legislation. At present, section 61 of the Human Rights Act 1993 criminalises only incitement of racial disharmony, leaving no equivalent protections for religion, gender, sexual orientation, disability, or other vulnerable identities. This narrow formulation fails to reflect the plural and multicultural reality of Aotearoa New Zealand.

The Royal Commission of Inquiry into the Christchurch Terrorist Attack (2020) recommended that these protections be expanded to ensure parity with comparable jurisdictions.[807] The proposal aligns with the approach in the United Kingdom, Australia, and the European Union, where hate-speech prohibitions recognise a broader range of protected grounds. From Waldron's perspective, such reform strengthens public dignity and equal assurance by signalling that all groups are entitled to protection from vilifying expression.

This thesis therefore recommends that any statutory duty-of-care framework be complemented by reform of the *Human Rights Act 1993*, either through amendment of section 61 or through dedicated online-safety legislation that explicitly extends protection to religion, gender, sexual orientation, and disability. Doing so would create coherence between offline and online

[807] Royal Commission of Inquiry Into The Terrorist Attack On Christchurch Mosques On 15 March 2019 "Part 9 - Social cohesion and embracing diversity" (2021) <https://christchurchattack.royalcommission.nz/the-report/part-9-social-cohesion-and-embracing-diversity/hate-crime-and-hate-speech> at 50.

protections, reinforce the values of equality and dignity under the *New Zealand Bill of Rights Act 1990*, and respond directly to the Royal Commission's call for stronger, inclusive safeguards.

6.5 Conclusion

The comparative analysis of the United Kingdom, Australia, and New Zealand demonstrates an international move from self-regulation toward legally enforceable duties of care for online platforms. The United Kingdom's Online Safety Act 2023 and Australia's Online Safety Act 2021 illustrate two different models for achieving the same goal: making social media platforms legally responsible for preventing foreseeable online harms.[808] Although New Zealand's current regime, centred on the Harmful Digital Communications Act 2015 and Human Rights Act 1993, addresses individual instances of harmful communication, it lacks a systemic framework that ensures proactive platform accountability.

In order to close this gap, New Zealand should adopt a statutory duty of care for large social media services, integrated into a dedicated Online Safety Act.[809] The duty should require platforms to take reasonable and proportionate steps to identify and mitigate foreseeable risks of online hate speech and related harms. This obligation would establish a legal baseline for platform accountability and move the regulatory approach from reactive complaint-handling to proactive harm prevention.

[808] *Online Safety Act 2023* (UK); *Online Safety Act* 2021 (Cth) Netsafe, above n 800.
[809] Ministry of Justice "Proposal against incitement of hatred and discrimination" (2021) <https://www.justice.govt.nz/assets/Documents/Publications/Incitement-Discussion-Document.pdf>.

An independent Online Safety Commission should be created to enforce this statutory duty. Operating independently from ministerial control, the Commission would oversee compliance by developing and approving industry codes, conducting regular audits of high-risk services, and issuing penalties for repeated or serious non-compliance. Alternatively, this role could be undertaken by an empowered version of Netsafe, provided its mandate is strengthened through statute and supported by stable public funding. The Online Safety Commission (or an equivalent regulator) would also coordinate with the Privacy Commissioner and Classification Office to ensure consistency and prevent regulatory overlap.

These reforms would transform New Zealand's online-safety architecture from a reactive complaint-driven model into a preventive system grounded in accountability, transparency, and public trust. The Online Safety Act would operate in parallel with the Harmful Digital Communications Act 2015, providing a systemic framework for platform governance, while the Human Rights Act 1993 should be amended to expand protected characteristics, aligning offline and online safeguards and reinforcing dignity under the New Zealand Bill of Rights Act 1990.

By adopting these reforms, New Zealand can strengthen protection against online hate speech and harmful content without undermining democratic participation or legitimate expression. This approach would position New Zealand as a leader in proportionate, rights-based digital regulation and ensure that the law keeps pace with the realities of online communication.

What these recommendations also do is translate the theoretical insights developed in Chapter 2 into a concrete regulatory design for New Zealand. Mill's harm principle underpins the preventive rationale for a statutory duty of care, while Meiklejohn's emphasis on informed

democratic participation supports the need for transparency and accountability. Waldron's focus on dignity and public assurance grounds the moral imperative for inclusive protection, and Breyer's proportionality framework ensures that regulation remains balanced and rights-consistent. Collectively, these perspectives justify a model of online-safety regulation that is both principled and practical, embedding freedom of expression and protection from harm within a single coherent framework.

Chapter 7: Conclusion

This thesis has examined how New Zealand has responded to online hate speech and harmful content in three main areas: law and policy, platform accountability, and technological design. It examined the Harmful Digital Communications Act 2015 together with related statutes. It also considered whether platform self-regulation is reliable, and by contrast, it drew on the EU's Digital Services Act and the UK's Online Safety Act to highlight options for reform in New Zealand, including the possible role of criminal law. These strands set the groundwork for analysing what each these approach achieves and where the gaps remain.

The chapters of this thesis build progressively towards the final argument. Chapter 2 set out the theoretical foundations. It does so by, drawing on free speech philosophy, international human rights law, and behavioural and regulatory theories such as the Online Disinhibition Effect and Lessig's modalities. Chapter 3 examined social media and algorithmic risks, showing how platform design and amplification mechanisms intensify the spread of online hate speech and harmful content, and why these dynamics challenge traditional legal regulation. Chapter 4 turned to New Zealand's domestic framework, analysing statutes such as the Harmful Digital Communications Act 2015 and the Human Rights Act 1993, and highlighting the reactive and fragmented character of current responses. Chapter 5 shifted to comparative

and transnational perspectives, focusing on the European Union, France, Germany, and hybrid initiatives like the Christchurch Call, and showing how proactive models contrast with New Zealand's reliance on voluntary self-regulation. Chapter 6 considered whether New Zealand should move towards a statutory duty of care, evaluating the models emerging in the United Kingdom and Australia and assessing how such an approach could be adapted to the New Zealand context.

At the centre of this study is the question of how New Zealand law should draw the line between protected and prohibited speech in the online environment. As outlined in Chapter 2, this thesis argues that the boundary cannot be fixed at the point of simple offensiveness. Rather, it should be located where expression causes demonstrable harm to dignity, inclusion, and the broader social fabric. This position reflects liberal philosophical principles while aligning with international human rights law, which recognises that freedom of expression carries corresponding duties and responsibilities, especially when speech undermines equality or human security.

This thesis has also engaged with key theoretical frameworks to better understand how these problems arise and why they persist. By applying the Online Disinhibition Effect (ODE) Theory and Regulation Theory, the research provides deeper insight into why harmful expression spreads so easily online and why current mechanisms struggle to contain it. The Online Disinhibition Effect (ODE) helps explain why online hate speech and harmful content grow so easily in digital spaces. Anonymity, low social accountability, and the absence of real-world consequences contribute to behaviour that would be less likely offline. While there is sometimes an assumption that existing laws addresses these patterns, the reality is different. New Zealand's legal frameworks, including the Harmful Digital Communications Act 2015,

tend to work only after harm has already occurred. They are reactive. This reactive approach leaves a gap. Social media platforms foster what can be called toxic disinhibition. Harmful communities can develop quickly, often without checks. Without a statutory duty on platforms to step in earlier, ODE-driven behaviours that fuel online hate speech and harmful content will continue unchecked.

Each chapter of this thesis contributes to addressing the central research question of how New Zealand can more effectively regulate online hate speech and harmful content. Chapter 2 outlined the theoretical and normative foundations of the study. Chapter 3 examined the role of social media design and algorithmic amplification in intensifying harm. Chapter 4 analysed the domestic legal framework, while Chapter 5 and Chapter 6 drew comparative insights from international models and proposed the statutory duty of care as a way forward.

The analysis across the preceding chapters shows each area of response (legal, institutional and technological) offering only a partial solution. The New Zealand framework remains fragmented and reactive, as discussed in Chapter 4, addressing harm only after it occurs and lacking clear platform duties. Chapter 5 demonstrated that self-regulation continues to dominate with voluntary measures such as Community Guidelines and the Christchurch Call limited by their dependence on platform goodwill and market incentives. Chapter 6 showed that while technological tools like AI moderation can assist, they remain prone to bias, overreach and inconsistent application. These findings confirm that a more integrated and proactive model is needed. Waldron's emphasis on dignity and inclusion helps explain why this integration matters. Hate speech and harmful content are not only about individual harm but also about the social fabric they corrode. When viewed this way, the case for proactive duties on platforms becomes stronger and aligns with international developments.

The need for a more integrated response can also be understood through Lessig's Regulation Theory, that shows how behaviour in digital spaces is not determined only by law but also by market forces, social norms and architecture. In New Zealand, the balance among these forces has tilted heavily toward voluntary compliance for example, through Community Guidelines and the Christchurch Call. These are well intentioned, but they are not enforceable, illustrating how self-regulation often fails to produce meaningful accountability. The architecture of social-media platforms reinforces this problem. Algorithms are designed to maximise engagement, which often means amplifying content that provokes strong emotion or outrage. As a result, online hate speech and harmful content are pushed forward rather than constrained, showing how design and profit incentives can undermine moderation efforts.

By contrast, overseas frameworks have shifted towards stronger and more proactive models of platform accountability. In the EU and the UK, legal duties require platforms to act before harm escalates, establishing proactive obligations rather than relying solely on reactive remedies. This thesis concludes that New Zealand should adopt a comparable statutory duty of care to hold platforms legally accountable for the online hate speech and harmful content they enable.

Fredman's model of substantive equality also helps explain why a stronger approach is needed. Legal responses to online hate speech and harmful content must move beyond content-based prohibitions and towards structural reform. Fredman's multidimensional framework (redressing disadvantage, enhancing voice, challenging stereotypes, and promoting institutional change) offers a compelling rationale for adopting a statutory duty of care. As discussed in Chapter 6, the proposed duty of care would operationalise these equity-based commitments into practice through enforceable obligations relating to risk-assessment, transparency and redress. As New Zealand considers reforms, the question is not only how to

regulate speech, but also how to design systems that uphold the values of dignity, equality, and democratic inclusion. This thesis contends that the statutory duty of care model, grounded in these commitments, presents the most promising path forward.

Across the analysis, this thesis has shown that platform accountability, technological solutions, and legal regulation each provide partial but incomplete responses to online hate speech and harmful content.

As discussed in Chapter 5, international developments confirm that stronger regulation is both possible and effective. Jurisdictions such as the European Union and the United Kingdom have already moved towards proactive duties, requiring platforms to assess and mitigate risks before harm escalates. These comparative models illustrate that voluntary compliance alone is no longer adequate, and they provide valuable guidance for New Zealand as it considers the design and scope of a statutory duty of care.

The central claim of this thesis is that a statutory duty of care, grounded in principles of dignity and substantive equality, offers the most effective way forward. Such a model would bring clarity to platform responsibilities, encourage proactive moderation, and support safer digital spaces. Reform must balance freedom of expression with protection from harm, but without stronger obligations, New Zealand's response to online hate speech and harmful content will remain fragmented, reactive, and inadequate in the digital age.

The contribution of this thesis lies in both theory and practice. It develops an integrated framework for understanding and addressing online hate speech and harmful content combining behavioural insights from the Online Disinhibition Effect, structural analysis from Lessig's Regulation Theory, and normative perspectives from Waldron's dignity-based approach and Fredman's model of substantive equality. By applying this framework to New

Zealand's fragmented and reactive legal response, the thesis provides the most comprehensive analysis to date of how existing statutes, platform self-regulation, and technological safeguards fall short. It also advances a normative and policy proposal: that New Zealand should adopt a statutory duty of care, supported by algorithmic transparency and grounded in principles of dignity, equality, and democratic inclusion. This combination of theoretical innovation, contextual analysis, and forward-looking recommendation represents the original contribution of the research.

Yet the path forward is not without challenges. Stronger platform accountability will attract political resistance both from the industry and from those concerned about freedom of expression. Technological change is also moving quickly, and regulation may struggle to keep pace with new forms of online harm. Jurisdictional limits mean that New Zealand cannot act alone but must work in step with international partners. These difficulties do not weaken the case for reform, but they emphasise the need for careful design, ongoing review, and a balanced approach.

While this thesis focuses on developing a statutory duty of care for regulating online hate speech and harmful content, it also points to areas that warrant further research. One area is the intersection between platform regulation and emerging technologies such as generative artificial intelligence, which will increasingly shape how harmful content is produced, amplified and moderated. Future research could explore how the duty of care might adapt to AI-driven environments, including questions of algorithmic transparency, automated decision-making and accountability for generative outputs.

Further work is also needed to examine how a duty of care framework could interact with the Te Tiriti o Waitangi and Māori data sovereignty principles, thereby ensuring that digital regulation in Aotearoa New Zealand reflects Indigenous rights and epistemologies.

Comparative research on Pacific and Asia-Pacific approaches to online harm would deepen regional context. Empirical studies that engage directly with affected communities and platform regulators would also deepen understanding of how regulatory obligations translate into real-world protection.

This research demonstrates that a statutory duty of care, grounded in dignity and equality, provides New Zealand with the best chance of building a future-proof framework to address online hate speech and harmful content in a rapidly changing digital environment. The task now is clear: New Zealand must move beyond reactive responses and commit to a framework that protects dignity, equality, and democratic participation in the digital age. This thesis shows that such a framework is both possible and necessary.

Bibliography

Cases

New Zealand

- *Ahsin v R* [2014] NZSC 153
- *Brooker v Police* [2007] NZSC 30.
- *Department of Internal Affairs v Young* [2004] DCR 231
- *Evans v R* [2008] DCR 199
- *R v Hansen* [2007] NZSC 7, [2007] 3 NZLR 1
- *Hooper v Gee* [2022] NZHC 1854
- *Hutton v R* [2018] NZCA 419
- *King - Ansell v Police* [1979] 2 NZLR 531 (CA)
- *Moonen v Film And Literature Board Of Review* [2000] 2 NZLR 9 (CA).
- *O'Neill v Malcouronne* [2021] NZHC 3027
- *Police v B* [2017] NZHC 526
- *R v Tarrant* [2020] NZHC 2192.
- *The King v Barker* [1924] NZLR 865
- *Wall v Fairfax New Zealand Ltd* [2018] NZHC 104
- *Wall v Fairfax New Zealand Ltd* [2017] NZHRRT 17

Australia

- *Coleman v Power* (2004) 220 CLR 1.

United Kingdom

- *Bourhill v Young* [1943] AC 92 (HL)
- *Donoghue v Stevenson* [1932] AC 562.
- *Handyside v United Kingdom* (1976) 1 EHRR 737 (ECHR)

North American

- *Cubby Inc v CompuServe Inc*, 776 F Supp 135 (SDNY, 1991).
- *Blumenthal v Drudge*, 992 F Supp 44 (DDC, 1998).
- *Fraley v. Facebook, Inc.* 830 F. Supp. 2d 785 (N.D. Cal. 2011).
- *Google Inc. v. Equustek Solutions Inc.* 2001 [2017] 1 SCR 824

- *McConnell v Federal Election Commission* 540 US 93 (2003),
- *Packingham v North Carolina* 137 S. Ct. 1730 (2017)
- *R v Oakes* [1986] 1 SCR 103
- *R.A.V. v City of St Paul* 505 US 377 (1992).
- *Religious Technology Center v Netcom On-Line Communication Services*, 907 F Supp 1361 (ND Cal, 1995) (copyright).
- *Snyder v Phelps* 562 US 443 (2011).
- *Zeran v America Online Inc*, 129 F 3d 327 (4th Cir, 1997).

Europe

- Case C 131-12 *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González* ECLI:EU:C:2014:317
- *Eva Glawischnig-Piesczek v Facebook Ireland Limited Case C-18/18* 2019 ECLI:EU:C:2019:821
- *Garaudy v France* (2003) V ECHR 383 (ECHR, App No 65831/01)
- Bundesgerichtshof [BGH] [Federal Court of Justice] Jul. 29, 2021, 154 Entscheidungen des Bundesgerichtshofes in Zivilsachen [BGHZ] 370, 371 (Ger.).

Legislation

New Zealand

- Broadcasting Act 1989
- Companies Act 1993
- Crimes Act 1961
- Film, Video, and Publications Classifications Act 1993
- Harmful Digital Communications Act 2015
- Human Rights Act 1993
- New Zealand Bill of Rights Act 1990
- Search and Surveillance Act 2012
- Summary Offences Act 1981
- Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2020

- Human Rights (Prohibition of Discrimination on Grounds of Gender Identity or Expression, and Variations of Sex Characteristics) Amendment Bill, Member's Bill 2023

Europe

- Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act, NetzDG) - Basic Information (2017) 2021 Bundesministerium der Justiz und für Verbraucherschutz.

- Charter of Fundamental Rights of The European Union 2012 European Union [2012] OJ 2012/C 326/02.

- Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law [2008] OJ L/328.
- Directive (EU) 2015/1535 Of The European Parliament And Of The Council 2015 [2015] OJ L 241/1.
- Directive (EU) 2015/1535.
- Directive 2000/31/EC.2021.
- Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') 2000 OJ L/178.
- German Criminal Code 1998.
- German Basic Law (Grundgesetz, GG) 1949.
- International Convention on the Elimination of All Forms of Racial Discrimination General comment No. 35, Combating racist hate speech CERD/C/GC/35 (26 September 2013).
- International Covenant on Civil and Political Rights (opened for signature 16 December 1066, entered into force 23 March 1976).
- International Covenant on Civil and Political Rights "General comment No. 34, Article 19: Freedoms of opinion and expression" CCPR/C/GC/34.
- Official Records of the General Assembly, Fifty ninth Session, Supplement No. 40, vol. I (A/59/40 (Vol. I)), annex III.

- United Nations General Assembly Report of the Special Rapporteur on the promotion and protection of the freedom of opinion and expression UN Doc A/74/486 (9 October 2019).
- Regulation 2022/2065 of the European Parliament and of the on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJEU L 277/1.
- Statute of the International Court of Justice.
- Treaty establishing the European Coal and Steel Community (18 April 1951).
- United Nations Human Rights *International Covenant on Civil and Political Rights* (23 March 1976).
- Vienna Convention on the Law of Treaties, United Nations, Treaty Series, vol. 1155, p. 331 (signed 23 May 1969, entered into force 27 January 1980).

Australia

- Online Safety Act 2021 (Australia).

Books and Chapters in Books

- Adam D. Thierer and Clyde Wayne Crews *Who rules the net?: Internet governance and jurisdiction* (Cato Institute, Washington, D.C, 2003)
- Aharon Barak *PART IV: Proportionality Evaluated*. In Proportionality, Vol. Series Number 2. United Kingdom: Cambridge University Press, 2012
- Alexander Brown and Adriana Sinclair *Hate Speech Frontiers: Exploring the Limits of the Ordinary and Legal Concepts* (Cambridge University Press, 2023)
- Alexander Meiklejohn, *Free Speech and Its Relation to Self-Government* (Harper & Brothers, 1948)
- Amin Mekacher, Max Falkenberg, Andrea Baronchelli "The systemic impact of deplatforming on social media" (2023) 2 PNAS Nexus
- Andreas Wiesand "Internet Content Suppression" *Culture and Human Rights: The Wroclaw Commentaries:* (De Gruyter, Berlin, Boston 2016)
- Aristotle - Translated by W. Rhys Roberts *Rhetoric* (350 B.C.E).

- AP Simester and WJ Brookbanks *Principles of Cirminal Law* (Thomson Reuters, Wellington, 2019)
- Brandon Reagen and others *Deep Learning for Computer Architects* (Springer International Publishing AG, Switzerland, 2022).
- Christopher Ram "Cybercrime" in Neil Boister and Robert J. Currie (ed) *Routledge Handbook of Transnational Criminal Law* (Routledge, London, 2014).
- David Harvey *internet.law.nz* (Fifth ed, LexisNexis NZ Limited, 2023).
- Dan Jerker B. Svantesson "Internet Jurisdiction And Intermediary Liability" in Giancarlo Frosio (ed) *Oxford Handbook of Online Intermediary Liability* (Oxford University Press, Oxford 2020).
- Dan Jerker B. Svantesson "5A New Jurisprudential Framework for Jurisdiction" *Solving the Internet Jurisdiction Puzzle* (Oxford University Press, 2017)
- David Bromell *Recent Developments in Other Selected Jurisdictions* (Springer Nature, Switzerland, 2022)
- E. Eugene Clark *Cyber law in Australia* (Kluwer Law International, Alphen aan den Rijn, The Netherlands, 2010) 318.
- Giancarlo Frosio Oxford Handbook of Online Intermediary Liability (Oxford University Press, Oxford, 2020).
- Hilary Gatti "*The Humanities as the Stronghold of Freedom" John Milton's Areopagitica and John Stuart Mill's On Liberty"* in Rens Bod, and others (eds) The Making of the Humanities (Amsterdam University Press, Amsterdam, 2019) 167
- Jaani Riordan, *The Liability of Internet Intermediaries* (Oxford University Press, Oxford, 2016)
- Jacinta Ruru and others The New Zealand Legal System : Structures and processes (6th ed, Lexis Nexis, Wellington, 2016).
- James Crawford *Brownlie's Principles of Public International Law* (Oxford University Press, 2012)524.
- Jennifer Jacobs Henderson "New Boundaries of Free Speech in Social Media" in Daxton R Stewart (ed) *Social media and the law : A guidebook for communication students and professionals* (2nd ed, Routledge, New York 2017).
- Jeremy Waldron *The Harm in Hate Speech* (Harvard University Press, London, 2012) Julia Hörnle "Head in the Clouds: The Clash between Territorial Sovereignty, Jurisdiction, and the Territorial Detachment of the Internet" *Internet Jurisdiction Law and Practice* (Oxford University Press, Oxford, 2021).

- Julie E. Cohen *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice* (Yale University Press, New Haven, 2012)
- John Stuart Mill *On Liberty*, edited by David Bromwich, and George Kateb (Yale University Press, 2003)
- Karl-Dieter Opp *Edwin H. Sutherland's Differential Association Theory* (1st ed., Routledge, 2020) at 138.
- Laurent Pernot *Epideictic Rhetoric in Ancient Greece* (University of Texas Press, Austin, 2015).
- Lawrence Lessig *Code: And Other Laws of Cyberspace* (Basic Books, New York, 1999).
- Leonie Rösner and Nicole C. Krämer "Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments" 2016 2 Social media + society.
- Michael Crotty *The Foundations of Social Research: Meaning and Perspective in the Research Process* (Sage Publications, London, 1998)
- Michael Geist, "The Equustek Effect: A Canadian Perspective on Global Takedown Orders In The Age Of The Internet" in Frosio *Oxford Handbook of Online Intermediary Liability* at 710.
- Neil Boister "The Concept of Transnational Criminal Law" in Neil Boister and Robert J. Currie (ed) *Routledge Handbook of Transnational Criminal Law* (Routledge, London, 2014)
- Nicole Stremlau and Iginio Gagliardone "Socio-legal approaches to online hate speech" in Naomi Creutzfeldt, Marc Mason, and Kirsten McConnachie (eds) Routledge Handbook of Socio-Legal Theory and Methods (Routledge, 2019) 385.
- Pierre Legrand *Paradoxically, Derrida: for a Comparative Legal Studies. In Derrida and Law* (1st ed., Routledge, 2009)
- Pablo Nicolás Terevinto and others "A Framework for OSN Performance Evaluation Studies" in Tansel Özyer and Reda Alhajj (eds) *Machine Learning Techniques for Online Social Networks* (Cham, Springer International Publishing, 2018).

- Richard Corbett and others European Parliament (John Harper Publishing, London, 2011).

- Roger S. Clark "Jurisdiction Over Transnational Crime" in Neil Boister and Robert J. Currie (ed) *Routledge Handbook of Transnational Criminal Law* (Routledge, London, 2014).
- Stephanie Tom Tong "Social Processes of Online Hate" in Diana E Forsythe and Ashley Marie Mehlenbacher (eds) *Digital Hate: The Global Convergence of Hate Online* (Routledge, London, 2023)
- Stephen Breyer Active Liberty: Interpreting Our Democratic Constitution (The Tanner Lectures on Human Values, Harvard University, 2004)
- Stephen Todd (ed) The Law of Torts in New Zealand (3rd ed, Wolters Kluwer Law International, 2020)
- Tarleton Gillespie "Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media" (2018) Yale University Press
- Tim Boone and others "Social Learning Theory Albert Bandura Englewood Cliffs, N.J.: Prentice-Hall, 1977. 247 pp. paperbound Group & Organisation Studies.

Journal Articles

- Alfred de Zayas "W.A. Schabas, Nowak's CCPR Commentary: U.N. Covenant on Civil and Political Rights" Neth Int Law Rev 67 at 558
- Amit M. Sachdeva "International Jurisdiction In Cyberspace: A Comparative Perspective" 2007 Computer and Telecommunications Law Review 245.
- Andrew Guess, Jonathan Nagler and Joshua Tucker, "Less than you think: Prevalence and predictors of fake news dissemination on Facebook" (2019) 5 Science Advances 4586.
- Andrew Geddis "The State of Freedom of Expression in New Zealand: An Admittedly Eclectic Overview" (2008) 11 Otago L Rev 657
- Andrew Murray The Regulation of Cyberspace (Routledge-Cavendish, Abingdon, 2007).
- Andrew D Murray "Mapping the Rule of Law for the Internet" in David Mangan and Laura E Gillies(eds) Legal Challenges of Social Media (Edward Elgar, Cheltenham, 2017) 13.
- Alexander Brown "What is Hate Speech? Part 1: The Myth of Hate" (2017) 36 Law and Philosophy.

- Agnieszka Pluta and others "Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain" (2023) 13 Scientific Reports.
- Ariadna Matamoros-Fernández and Johan Farkas "Racism, Hate Speech, and Social Media: A Systematic Review and Critique" (2021) 22 Television & New Media.
- Asma Vranaki, 'Review of The Regulation of Cyberspace: Control in the Online Environment by Andrew D Murray' (2008) 1 Journal of Information, Law & Technology.
- Barbara Perry & Patrik Olsson "Cyberhate: The globalization of hate" (2009) 18 Information & Communications Technology Law 185.
- Brian O'Shea "A New Method to Address Cyberbullying in the United States: The Application of a Notice-and-Takedown Model as a Restriction on Cyberbullying Speech Notes" 2017 69 Fed. Comm. L.J. 121.
- Caio C.V. Machado and Thais Helena Aguiar "Emerging Regulations on Content Moderation and Misinformation Policies of Online Media Platforms: Accommodating the Duty of Care into Intermediary Liability Models." (2023) Business and Human Rights Journal 8
- Damian Tambini "The differentiated duty of care: a response to the Online Harms White Paper" 2019 11 Journal of Media Law 359.
- Dan Jerker B. Svantesson "Lagom jurisdiction" - What drinking etiquette can teach us about internet jurisdiction and Google France" 2018 Masaryk University Journal of Law and Technology 29.
- Dan Svantesson "An Introduction to Jurisdictional Issues in Cyberspace" (2004) Journal of Law, Information and Science 50 <http://classic.austlii.edu.au/au/journals/JlLawInfoSci/2004/3.html>.
- Dan Jerker B. Svantesson "A New Jurisprudential Framework For Jurisdiction: Beyond The Harvard Draft" (2015) 109 AJIL Unbound 69.
- Dan Jerker B. Svantesson "Jurisdictional issues and the internet - a brief overview 2.0" (2018) 34 Computer Law & Security Review 715.
- Dan Jerker B. Svantesson "Delineating the Reach of Internet Intermediaries' Content Blocking - ccTLD Blocking, Strict Geo-Location Blocking or a Country Lens Approach" 2014 11 SCRIPTed.
- Dan Valeriu Voineo, "Taking Over Twitter - Balancing Free Speech And Content Moderation" (2022) 8 Annals of the University of Craiova for Journalism, Communication and Management 139

- David Kay cited in Evelyn Mary Aswad (2020) 609 W&L L Rev 618
- David La Barbera, Eddy Maddalena, Michael Soprano, Kevin Roitero, Gianluca Demartini, Davide Ceolin, Damiano Spina and Stefano Mizzaro “Crowdsourced Fact-checking: Does It Actually Work?” (2024) 61 Information Processing & Management < https://doi.org/10.1016/j.ipm.2024.103792>
- Didier Musiedlak "The Metamorphoses of Mussolini’s Body" 2018 20 Journal of Genocide Research 236.
- Dominic Abrams “Processes of prejudice: Theory, evidence and intervention. Equality and Human Rights Commission Research report” (2010) 56 Center for the study of Group Processes, University of Kent.
- Dylan Asafo “The Western Legal Roots of the Christchurch Mosque Shootings: A Colonial Critique of New Zealand’s Legal Framework on Racist Hate Speech” (2021) 12 *UC Irvine L Rev* 101
- Erika Szyszczak “Antidiscrimination Law in the European Community” (2009) 32 Fordham International Law Journal 623
- Evan Brody, Spencer P. Greenhalgh, and Mehroz Sajjad “Gayservatives on Gab: LGBTQ+ Communities and Far Right Social Media” 8 SM+S
- Evelyn Douek "Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility" (2021) 21 North Carolina Journal of Law & Technology 1.
- Evelyn Douek "Governing Online Speech: From "Posts-As-Trumps" To Proportionality And Probability" (2021) 121 Columbia Law Review 759.
- Giulio Corsi “Evaluating Twitter’s algorithmic amplification of low-credibility content: an observational study” (2024) 13 EPJ Data Science
- "Hitler On Propaganda" (1951) 37 Quarterly Journal of Speech 440.
- Hunt Allcott and Matthew Gentzkow “Social Media and Fake News in the 2016 Election” (2017) 31 The Journal of Economic Perspectives 211.
- Jaakko Husa, “About the Methodology of Comparative Law: Some Comments Concerning the Wonderland” (2015) 59 Revue Internationale de Droit Comparé 504
- James Hawdon, Atte Oksanen, and Pekka Räsänen “Exposure to Online Hate in Four Nations: A Cross-National Consideration” (2017) 38 Deviant Behavior 254 at 265.
- Joel R. Reidenberg “Lex Informatica: The Formulation of Information Policy Rules through Technology” , 76 Tex. L. Rev. 553 (1997-1998) 28
- Joel Postman SocialCorp : social media goes corporate (New Riders, Berkeley, CA, 2009)

- Jonathan A. Obar and Anne Oeldorf-Hirsch "The Clickwrap: A Political Economic Mechanism for Manufacturing Consent on Social Media" 2018 4 Social Media + Society at 3.
- John Suler, “The Online Disinhibition Effect” (2005) 2(2) International Journal of Applied Psychoanalytic Studies 184
- John R. Suler Psychology of the Digital Age: Humans Become Electric (Cambridge University Press, Cambridge, 2015).
- Janiesch C, Zschech P and Heinrich K, 'Machine learning and deep learning' (2021) 31 Electronic Markets 685.
- Katerina Linos "How to select and develop international law case studies: lessons from comparative law and comparative politics" (2015) 109 American Journal of International Law 475.
- Kate Crawford and Tarleton Gillespie "What is a flag for? Social media reporting tools and the vocabulary of complaint" (2016) 18 New Media & Society 410.
- Katharine Gelber and Luke McNamara “Evidencing the harms of hate speech” (2015) Social Identities 324
- Karen M. Douglas, Jan-Willem van Prooijen and Robbie M. Sutton, ‘Is the label ‘conspiracy theory’ a cause or a consequence of disbelief in alternative narratives?’ British Journal of Psychology, 113(3) 2022.
- Karine Barzilai-Nahon "Toward a theory of network gatekeeping: A framework for exploring information control" 2008 59 Journal of the American Society for Information Science and Technology 1493 at 1494.
- Karine Perset “The Economic and Social Role of Internet Intermediaries” OECD Digital Economy Papers No 171 (2010)
- Kylie Pappalardo and Nicholas “The Liability of Australian Online Intermediaries” (2018) 40 Sydney Law Review 469.
- Kenneth Grad and Amanda Turnbull “Harmful Speech and The COVID-19 Penumbra” (23 August 2021) 19 Canadian Journal of Law and Technology 41.
- Kristin Henrard, “EU Law’s Half-Hearted Protection of Religious Minorities” (2021) 12 Religions 830
- Laurent Pech, “The Rule of Law as a Constitutional Principle of the European Union” (2010) 6 EuConst 359
- Lawrence Lessig "The New Chicago School" (1998) 27 The Journal of Legal Studies 661.

- Lawrence Lessig "The Regulation of Social Meaning" (1995) 62 The University of Chicago Law Review 994.
- Lorna Woods "The duty of care in the Online Harms White Paper" 2019 11 Journal of Media Law.
- Lea Bader and Jochen Bender “What is “fake news” and “hate speech” and how do they work in practice?” Central and Eastern European EDem and EGov Days 342 (March 2022) 17.
- Matteo Cinelli and others “Dynamics of online hate and misinformation” (2021) 11(1) Scientific Reports
- Michael O’Flaherty “Freedom of Expression: Article 19 of the International Covenant on Civil and Political Rights and the Human Rights Committee’s General Comment No 34” (2012) HRLR 627
- Mathew Binny and others "Thou Shalt Not Hate: Countering Online Hate Speech" (2019) 13 Proceedings of the International AAAI Conference on Web and Social Media Full Papers.
- Nicolas Suzor, “Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms” (2018) 4(3) Social Media + Society 1.
- Nehaluddin Ahmad and Norulaziemah Zulkiffle "Jurisdiction Issues in Cyberspace: An Overview in Respect of Brunei and Malaysia Compared to The United States' System" 2022 Journal of Southeast Asian Research Article ID 382477.
- Noam Lapidot-Lefler and Azy Barak "Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition" (2012) 28
- Paul Benjamin Lowry and others “Why Do Adults Engage in Cyberbullying on Social Media? An Integration of Online Disinhibition and Deindividuation Effects with the Social Structure and Social Learning Model” 2016 27(4) Information Systems Research.
- Paul De Hert and Eugenio Mantovani, “Review of The Regulation of Cyberspace” (2008) 2(1) Studies in Ethics, Law and Technology
- Rachel Sue Yin Tan "Disabling access to illegal online content by way of takedowns" [2021] NZLJ 341.
- Rachel Sue Yin Tan, ‘Social Media Platforms - Duty of Care’ (2022) 36 Australasian Parliamentary Review 143.

- Rachel Sue Yin Tan “Legislative Strategies to Tackle Misinformation and Disinformation: Lessons from Global Jurisdictions” (2023) 38 Australasian Parliamentary Review 231.
- Richard Rogers "Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media:" (2020) 35 European Journal of Communication 201.
- Robert C Post, ‘Review of Freedom of Speech by Eric Barendt’ (1988) 36(1) Am J Comp L 174
- Robert Gorwa, “The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content” (2019) 8(2) Internet Policy Review
- Robert Mark Simpson "Dignity, Harm, and Hate Speech" (2013) 32 Law and Philosophy 701 at 727.
- Rosara Joseph "Inherent jurisdiction and inherent powers in New Zealand" 2005 220 Canterbury Law Review.
- Rebecca Zipursky "Nuts About NETZ: The Network Enforcement Act and Freedom of Expression- " (2019) 42 Fordham International Law Journal 1325.
- Reza Farahbakhsh, Marzieh Mozafari, and Noël Crespi "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media" (paper presented to the Eighth International Conference on Complex Networks and Their Applications, France , 2019).
- Ruben Enikolopov and others "Social Media and Protest Participation: Evidence From Russia" (2020) 88 Econometrica 1479.
- Sandra Fredman “Substantive Equality Revisited” (2016) 14(3) International Journal of Constitutional Law 712
- Sebastian Wachs and Michelle F. Wright "Associations between Bystanders and Perpetrators of Online Hate: The Moderating Role of Toxic Online Disinhibition" (2018) 15 International Journal of Environmental Research and Public Health
- Sally Broughton Micova and Alexandre de Streel Digital Services Act - deepening the internal market and clarifying responsibilities for digital services (Centre on Regulation in Europe, 2020).
- Shahram Akbarzadeh, Amin Naeni, Ihsan Yilmaz, and Galib Bashirov “Cyber Surveillance and Digital Authoritarianism in Iran” (Vol 14, Issue 3, Global Policy, March 2024)

- Stephanie Panzic “The Harmful Digital Communications Act 2015: A Deficiency in Defining Harm and an Unintended Limitation on Freedom of Expression” (2015) 21 Auckland U L Rev 218
- Ulrike Klinger “Digital Democracy and Public Discourse: Dissonant, Disrupted and Unedited?” (Vol 69, No 26, Canadian International Council, 2021)
- W.J. Kamba, ‘Comparative Law: A Theoretical Framework’ (1974) 23(3) International and Comparative Law Quarterly 485
- Wolfgang Schulz "Regulating Intermediaries to Protect Privacy Online - The Case of the German NetzDG" 2018.

Parliamentary and Government Materials

New Zealand

- (15 August 2019) Written Questions 29678 (Address in Reply, Tim Macindoe).
- Department of the Prime Minister and Cabinet- Te Tari O Te Pirimia Me Te Komiti Matua *Cabinet Manual* (2017).
- Henry Talbot and Alali Nusiebah “The Edge of the Infodemic: Challenging Misinformation in Aotearoa” (30 June 2021) Te Mana Whakaatu <https://www.classificationoffice.govt.nz/news/news-items/the-edge-of-the-infodemic/>
- Human Rights (Prohibition of Discrimination on Grounds of Gender Identity or Expression, and Variations of Sex Characteristics) Amendment Bill, Member’s Bill, 275-1
- Jacinda Ardern "Significant progress made on eliminating terrorist content online" (press release, 24 September 2019).
- Justice and Electoral Committee *Harmful Digital Communications Bill: Report of the Committee* (March 2014)
- Labour 2020 "Our Manifesto to Keep New Zealand Moving" <https://www.labour.org.nz/policy>.

- Law Commission *Harmful Digital Communications: The adequacy of the current sanctions and remedies* (Ministerial Briefing Paper, August 2012, Wellington) (Law Commission Paper).
- Law Commission Hate Crime Law Reform (2024) <https://www.lawcom.govt.nz/our-work/hate-crime/>.
- Law Commission *A New Zealand Guide to International Law and its Sources* (NZLC R34, 1996).
- Netsafe "Report - Netsafe - Providing free online safety advice in New Zealand" (2022) <https://www.netsafe.org.nz/reportanincident/>.
- Netsafe "Aotearoa New Zealand Code of Practice for Online Safety and Harms draft - Netsafe - Providing free online safety advice in New Zealand" (2021) <https://www.netsafe.org.nz/aotearoa-new-zealand-code-of-practice-for-online-safety-and-harms-draft/>.
- NZ Human Rights Commission "Meng Foon: Covid-19 coronavirus fear no excuse for racism" (2020) <https://www.hrc.co.nz/news/meng-foon-covid-19-coronavirus-fear-no-excuse-racism/>.
- The Classification Office - Te Mana Whakaatu "Christchurch Mosque Attack Livestream" (2019).
- Te Kāwanatanga o Aotearoa New Zealand Government "Christchurch Call" <https://www.christchurchcall.com/call.html>.
- William Young and Jacqui Caine *Royal Commission of Inquiry into The Terrorist Attack on Christchurch Mosques On 15 March 2019* (November 2020).

Europe

- Alexandre De Streel and others *Online Platforms 'Moderation of Illegal Content Online* (Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, PE 652.718, 2020)
- Andrea Bertolini and others "Liability of Online Platforms" February 2021 European Parliament Research Service (EPRS), February 2021).

- Bundesministerium der Justiz und für Verbraucherschutz "FAQ: Act to Improve Enforcement of the Law in Social Networks, 2017" (2021).

- European Commission "European Commission and IT Companies announce Code of Conduct on Illegal Online Hate Speech" (2016) Press Release .
- European Commission "The Code of conduct on countering illegal hate speech online" (2021).
- Consolidated version of the Treaty on European Union (26 October 2010) *Official Journal of the European Union C 326/15*
- Committee of Experts on Combating Hate Speech (ADI/MSI-DIS) "Background document" Combating Hate Speech (ADI/MSI-DIS) (25 May 2020) Council of Europe.
- Proposal for a Regulation of The European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC.2021
- European Council - Council of the European Union <https://www.consilium.europa.eu/en/council-eu/>.
- European Union "The history of the European Union" < https://european-union.europa.eu/principles-countries-history/history-eu_en>.
- European Commission "Adopting EU Law" (2021) <https://ec.europa.eu/info/law/law-making-process/adopting-eu-law_en>.
- European Commission "The European Commission's priorities" <https://ec.europa.eu/info/strategy_en>.
- European Commission "Digital Services Act - deepening the internal market and clarifying responsibilities for digital services" (2021) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12417-Digital-Services-Act-deepening-the-internal-market-and-clarifying-responsibilities-for-digital-services>.
- European Commission "Relations with national parliaments" (2016) <https://commission.europa.eu/law/law-making-process/adopting-eu-law/relations-national-parliaments_en >
- European Commission "The EU Single Market" < https://single-market-economy.ec.europa.eu/single-market_en>
- European Parliament Information Office <https://europarlamentti.info/en/values-and-objectives/objectives/>.
- European Parliament "Infringement procedure" < https://commission.europa.eu/law/application-eu-law/implementing-eu-law/infringement-procedure_en>

- European Parliament "Supervisory Powers" (2020) <https://www.europarl.europa.eu/about-parliament/en/powers-and-procedures/supervisory-powers>.
- European Parliament “Hate speech and hate crime in the EU and the evaluation of online content regulation approaches” (July 2020) <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf>
- European Union "Court of Justice of the European Union (CJEU)" <https://europa.eu/european-union/about-eu/institutions-bodies/court-justice_en>.
- European Commission "Combined Evaluation Roadmap/Inception Impact Assessment Digital Services Act package: deepening the Internal Market and clarifying responsibilities for digital services” (June 2020)

- European Commission “2022 Strengthened Code of Practice on Disinformation” <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation>

United Kingdom

- UK Parliament “Parliamentary Bills - Online Safety Bills” (2023).
- House of Lords, Select Committee on Communications, Parliament of United Kingdom, *Regulating in A Digital World*, 2nd Report, Session 2017-19.
- Department for Science, Innovation and Technology “Centre for data Ethics and Innovation”
- Department for Digital, Culture, Media & Sport “Statutory guidance - Code of Practice for providers of online social media platforms” (12 April 2019).
- House of Lords, Select Committee on Communications, Parliament of United Kingdom, Regulating in A Digital World, 2nd Report, Session 2017-19.

Australia

- House of Representatives Select Committee on Social Media and Online Safety *Social Media and Online Safety* (March 2022).

- Commonwealth of Australia Government Response and Implementation Roadmap for the Digital Platforms Inquiry "*Regulating in the Digital Age"* (2019)
- eSafety Commissioner "Online Safety Act 2021 Fact sheet" (2021).
- Explanatory Memorandum House of Representatives "Social Media (Anti-Trolling) Bill 2022" (2022).
- Paula Pyburne and Rhonda Jolly *Australian Governments and dilemmas in filtering the Internet: juggling freedoms against potential for harm - Parliament of Australia* (8 August 2014).

Reports

- Alexandre de Streel and others Online Platforms' Moderation of Illegal Content Online - Law, Practices and Options for Reform (European Parliament's committee on Internal Market and Consumer Protection, Scientific Foresight Unit (STOA), PE 656.318, February 2021).
- Ciarán O'Connor *Hatescape: An In-Depth Analysis of Extremism and Hate Speech on TikTok* (Institute for Strategic Dialogue, 2021).
- Committee on the Films, Videos and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2021.
- Dan Svantesson *Internet Jurisdiction Global Status Report 2019: Key Findings* (2019).
- David Livingstone Paul Cornish, Dave Clemente and Claire Yorke Cyber Security and the UK's Critical National Infrastructure (A Chatham House Report, September 2011).
- European Commission Combined Evaluation Roadmap/Inception Impact Assessment Digital Services Act package: deepening the Internal Market and clarifying responsibilities for digital services Ref. Ares(2020)2877686 - 04/06/2020

- Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2020 (select committee report).
- Heidi Tworek and Paddy Leerssen *An Analysis of Germany's NetzDG Law* (Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, 15 May 2019).

- Human Rights Committee *General Comment No 34, Article 19: Freedoms of opinion and expression* CCPR/C/GC/34 (12 September 2011).
- Iginio Gagliardone and others *Countering online hate speech* (UNESCO, 2015).
- International Convention on the Elimination of All Forms of Racial Discrimination *General Recommendation No. 35 Combating Racist Hate Speec*h CERD/C/GC/35 (26 September 2013)
- Labour Party "Labour's 2020 Election Manifesto" (13 October 2020) <https://www.labour.org.nz/news-labour_2020_manifesto>.
- Law Commission "Ia Tangata, A review of the protections in the Human Rights Act 1993 for people who are transgender, people who are non-binary and people with innate variations of sex characteristics" (2025) <https://www.lawcom.govt.nz/our-work/ia-tangata/tab/report>
- New Zealand Human Rights Commission "Korero Whakamauahara-Hate Speech" 2019.
- Royal Commission of Inquiry Into The Terrorist Attack On Christchurch Mosques On 15 March 2019.
- Secretary of State of Digital, Culture, Media and Sport, *'Consultation Outcome - Online Harms White Paper: Full government response to the consultation'.* <https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response.>
- Susan S. Silbey "After Legal Consciousness." Annual Review of Law and Social Science 1, no. 1 (2005): 323-68. < doi:10.1146/annurev.lawsocsci.1.041604.115938.>
- The Classification Office - Te Mana Whakaatu "Christchurch Mosque Attack Livestream" (2019) <https://www.classificationoffice.govt.nz/news/featured-classification-decisions/christchurch-mosque-attack-livestream/>.
- The Classification Office - Te Mana Whakaatu "Christchurch Mosque Attack Livestream" (2019) <https://www.classificationoffice.govt.nz/news/featured-classification-decisions/christchurch-mosque-attack-livestream/>.
- The Office of the Governor-General "New Zealand's Constitution" (2020) <https://gg.govt.nz/office-governor-general/roles-and-functions-governor-general/constitutional-role/constitution/constitution>.
- Te Tari Taiwhenua - Department of Internal Affairs "Minimising Harm - Maximising Benefit: The Department of Internal Affairs Approach to Compliance & Enforcement 2012" (2012) <https://www.dia.govt.nz/Minimising-Harm-Maximising-Benefit>.

- United Nations "United Nations Detailed Guidance on Implementation for United Nations Field Presences" (2020) < https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf>
- United Nations "UN chief appeals for global action against coronavirus-fueled hate speech" (8 May 2020) United Nations Global Perspective Human Stories .<https://news.un.org/en/story/2020/05/1063542>.

- U.S. Department of Labor *Issues in Labor Statistics, Bureau of Labor Statistics* (March 1999).
- Wilson Richard Ashby "HATE: Why We Should Resist it with Free Speech, Not Censorship by Nadine Strossen (review)" 2019 41 Human Rights Quarterly 213.

Dissertations

- Emily Laidlaw "Internet Gatekeepers, Human Rights and Corporate Social Responsibilities" (Doctor of Philosophy thesis London School of Economics and Political Science, 2012).
- Graham Edward Geddes "Keyboard Warriors : The Production of Islamophobic Identity and an Extreme Worldview within an Online Political Community" (Doctor of Philosophy thesis University of York, 2014).
- Laura C. Rodgriguez Rengifo "Liability of Internet Intermediaries: Participative Networking Platforms and Harmful Content" (LLM University of Wellington, 2016).
- Jack Edmond "Potential responses to the threat of 'fake news' in a digitalised media environment" (LLB (Hons) Dissertation, University of Otago, 2018).

Internet resources

- ACT New Zealand "New bill will protect freedom of expression" (2019) ACT <https://www.act.org.nz/new_bill_will_protect_freedom_of_expression>.
- Alvin Chang "The Facebook and Cambridge Analytica scandal, explained with a simple diagram. A visual of how it all fits together. They're now shutting down." (2018)

<https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram>.

- Amélie Heldt "Germany is amending its online speech act NetzDG... but not only that" (6 April 2020) Leibniz Institute for Media Research, Hans-Bredow-Institut < https://policyreview.info/articles/news/germany-amending-its-online-speech-act-netzdg-not-only/1464 >
- Arielle Pardes "To Clean Up Comments Let AI Tell Users Their Words Are Trash" (2020) Wired <https://www.wired.com/story/comments-section-clean-up-let-ai-tell-users-words-trash/ >.
- Asha Barbaschow "Facebook, Google, Microsoft, TikTok, and Twitter adopt Aussie misinformation code | ZDNet" (2021) <https://www.zdnet.com/article/facebook-google-microsoft-tiktok-and-twitter-adopt-aussie-misinformation-code/>.
- Aleem Maqbool "Black Lives Matter: From social media post to global movement" (10 July 2020) BBC News < https://www.bbc.com/news/world-us-canada-53273381>
- Angela Boundy "2019 online hate speech insights" (12 December 2019) Netsafe - Online Safety Help and Advice for New Zealanders <https://www.netsafe.org.nz/2019-online-hate-speech-insights/>.
- Alex Mann and others "Christchurch shooting accused Brenton Tarrant supports Australian far-right figure Blair Cottrell" (2019) <https://www.abc.net.au/news/2019-03-23/christchurch-shooting-accused-praised-blair-cottrell/10930632>.
- Andrew Tobin and Jon Erbacher "The Israel Folau saga: A simple case of failure to comply with an employment contract or is there more?" (20 May 2019) < https://www.hopgoodganim.com.au/news-insights/the-israel-folau-saga-a-simple-case-of-failure-to-comply-with-an-employment-contract-or-is-there-more/>
- Andrew Little "Andrew Little: Hate speech threatens our right to freedom of speech" (27 April 2019) <https://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=12225871>.
- Association for Progressive Communications "Frequently asked questions on internet intermediary liability | Association for Progressive Communications" (2020) <https://www.apc.org/en/pubs/apc%E2%80%99s-frequently-asked-questions-internet-intermed>.
- Ben Knight "Germany implements new internet hate speech crackdown" DW.COM (1 January 2018) < https://www.dw.com/en/germany-implements-new-internet-hate-speech-crackdown/a-41991590>.

- Business Standard "Australia's social media anti-trolling bill raises alarm for tech giants" (2022) <https://www.business-standard.com/article/international/australia-s-social-media-anti-trolling-bill-raises-alarm-for-tech-giants-122030700244_1.html >.
- Brian Fung "Facebook's Oversight Board is finally hearing cases, two years after it was first announced" (2021) CNN <https://edition.cnn.com/2020/10/22/tech/facebook-oversight-board/index.html>.
- Brooke Tanner "EU Code of Practice on Disinformation" <https://www.brookings.edu/blog/techtank/2022/08/05/eu-code-of-practice-on-disinformation/>
- Case C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited Press and Information* (3 October 2019) Global Freedom of Expression < https://globalfreedomofexpression.columbia.edu/cases/glawischnig-piesczek-v-facebook-ireland-limited/ >
- Case decision 2021-003-FB-UA 2021, Oversight Board Upholds Facebook Decision In Armenians In Azerbaijan Case (2021) <https://www.oversightboard.com/news/436612660860568-oversight-board-upholds-facebook-decision-case-2020-003-fb-ua/ >
- Center for Countering Digital Hate "The Disinformation Dozen - Why platforms must act on twelve leading online anti-vaxxers" (24 March 2021) <https://counterhate.com/research/the-disinformation-dozen>
- Craig Silverman "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook" (2016) Buzzfeed News <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook>
- Crystal Wu "TikTok algorithm promoted anti-Semitic death camp video" <https://www.newshub.co.nz/home/world/2020/07/tiktok-algorithm-promoted-anti-semitic-death-camp-video.html?utm_source=dlvr.it&utm_medium=twitter>
- Christopher Hughes "Most active social media networks New Zealand 2018" (2019) <https://www.statista.com/statistics/681840/new-zealand-most-popular-social-media-networks/>.
- "Critics say Twitter treats hate speech as being 'public interest'" (16 October 2019) Al Jazeera < https://www.aljazeera.com/ajimpact/critics-twitter-treats-hate-speech-public-interest-191016225423140.html >.

- Christopher McFadden "A Brief History of Facebook, Its Major Milestones" (2020) <https://interestingengineering.com/history-of-facebook>.
- Cynthia Vinney “Sutherland’s Differential Association Theory Explained” (2019) <https://www.thoughtco.com/differential-association-theory-4689191>.
- Chris Köver and Markus Reuter “TikTok curbed reach for people with disabilities” (2019) <https://netzpolitik.org/2019/discrimination-tiktok-curbed-reach-for-people-with-disabilities/>
- Chloe Hadavas "France’s New Online Hate Speech Law Is Fundamentally Flawed" (2020) <https://slate.com/technology/2020/05/france-hate-speech-law-lutte-contre-haine-sur-internet.html>.
- Christopher McFadden "A Brief History of Facebook, Its Major Milestones" (29 March 2023) Interesting Engineering <https://interestingengineering.com/culture/history-of-facebook>
- Christoph Schmon and Paige Collings “The Adoption of the EU’s Digital Services Act: A Landmark Year for Platform Regulation 2022 in Review” (26 December 2022) <https://www.eff.org/deeplinks/2022/12/adoption-eus-digital-services-act-landmark-year-platform-regulation-2022-year>
- Courtnet Lawton “TikTok ban in the United Sates: A necessary precaution or a misstep” (29 Jan 2025) < https://policyreview.info/articles/news/tiktok-ban-united-states-necessary-precaution-or-misstep/1822>
- Columbia Global Freedom of Expression "Google Spain SL v. Agencia Española de Protección de Datos" <https://globalfreedomofexpression.columbia.edu/cases/google-spain-sl-v-agencia-espanola-de-proteccion-de-datos-aepd/>.
- Columbia Global Freedom of Expression "The Case on Facebook’s Terms of Service" (2021) <https://globalfreedomofexpression.columbia.edu/cases/the-case-on-facebooks-terms-of-service/>.
- Chris Köver and Markus Reuter “TikTok curbed reach for people with disabilities” (2019) <https://netzpolitik.org/2019/discrimination-tiktok-curbed-reach-for-people-with-disabilities/>
- Dan Svantesson "Supreme Court of Canada challenges the idea of state sovereignty | OUPblog" 2017 < https://blog.oup.com/2017/08/supreme-court-canada-state-sovereignty/ >
- Daphne Keller "The Center for Internet and Society" (2021) <https://cyberlaw.stanford.edu/about/people/daphne-keller>.

- Daphne Keller "Broad Consequences of a Systemic Duty of Care for Platforms" (1 June 2020) The Center for Internet Society <https://cyberlaw.stanford.edu/blog/2020/06/broad-consequences-systemic-duty-care-platforms>.
- David Seymour and Andrew Little "Freedom of speech: Do we need to update our Human Rights Act?" (2019) <https://www.stuff.co.nz/national/politics/opinion/113785976/freedom-of-speech-do-we-need-to-update-our-human-rights-act>.
- Deepa Christopher “India - Chinese Apps banned as border tensions rise” (8 July 2020) Linklaters <https://www.linklaters.com/en/insights/blogs/digilinks/2020/july/india---chinese-apps-banned-as-border-tensions-rise>
- Derek Cheng "Christchurch Call update: Social media giants join forces to fight extremism" New Zealand Herald (online ed, 23 September 2019).
- Deanna Ritchie "Facebook/Meta - celebrating 20 years of astounding innocation and conflicting emotions" (2024) Readwrite.com <https://readwrite.com/facebook-meta-celebrating-20-years-of-astounding-innovation-and-conflicting-emotions/>.
- Devin Powell "Brains Make Decisions the Way Alan Turing Cracked Codes" (2015) Smithsonian Magazine <https://www.smithsonianmag.com/science-nature/brains-make-decisions-way-alan-turing-cracked-codes-180954212/>.
- Dr. Michael J. Garbade "A Simple Introduction to Natural Language Processing" (2018) Becoming Human <https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32>.
- Dr. Katarina Ristic Blog #23: From Norway to New Zealand: How a Serbian Internet Meme inspired radical right terrorists worldwide (7 September 2020) < https://www.uni-leipzig.de/newsdetail/artikel/blog-from-norway-to-new-zealand-how-a-serbian-internet-meme-inspired-radical-right-terrorists-wor >
- Digby Werthmuller “Anti-mandate protesters convoy on both North and South Islands” (7 July 2022) 1News
- Digital Industry Group “Australian Code of Practice on Disinformation and Misinformation” (2021) <https://digi.org.au/disinformation-code/>
- Digital Photography DP Review "American Society of Media Photographers warns about new Facebook T& Cs" (8 September 2013) <https://www.dpreview.com/articles/3293203654/american-society-of-media-photographers-warns-about-new-facebook-t-cs >.

- Deutsche Welle (www.dw.com) "German satirist Jan Böhmermann sues Angela Merkel over Erdogan poem remark | DW | 02.04.2019" (2021) <https://www.dw.com/en/german-satirist-jan-böhmermann-sues-angela-merkel-over-erdogan-poem-remark/a-48158329>.
- Ed Burns "Definition: Machine Learning" (September 2023) Search Enterprise AI <https://searchenterpriseai.techtarget.com/definition/machine-learning-ML>.
- Edgar Pacheco and Neil Melhuish “2019 online hate speech insights', Netsafe - Online Safety Help and Advice for New Zealanders” (12 December 2019) Netsafe <https://www.netsafe.org.nz/2019-online-hate-speech-insights>
- Edgar Pacheco and Neil Melhuish "Annual Population Survey 2017" (2018) Netsafe - Online Safety Help and Advice for New Zealanders <https://www.netsafe.org.nz/annual-population-survey-2017/>.
- EuroISPA is the globally the largest association of internet service providers. EuroISPA "Liability of Intermediaries EuroISPA Recap of Past Event: Liability of Intermediaries” (podcast, 29 September 2021) EuroISPA < https://www.euroispa.org/2021/10/recap-of-past-event-liability-of-intermediaries/>.
- Eva Corlett and Tess McClure “New Zealand police clash with anti-vaccine protesters at parliament, over 120 arrested” (10 February 2022) *The Guardian* <https://www.theguardian.com/world/2022/feb/10/new-zealand-police-clash-with-anti-vaccine-protesters-during-eviction-operation>
- Eva Corlett “Fire and clashes break out at New Zealand parliament as police move in to clear protest” (2 March 2022) *The Guardian* < https://www.theguardian.com/world/2022/mar/02/police-move-to-clear-new-zealand-protests-as-maori-king-calls-for-end-to-occupation >
- Facebook <https://www.facebook.com>.
- Facebook "Terms of Service Section 3.2" (2025) <https://www.facebook.com/terms.php >.
- Facebook "What are Facebook Products? " <https://www.facebook.com/help/1561485474074139?ref=tos>.
- Facebook "Community Standards - Objectionable Content" <https://www.facebook.com/communitystandards/objectionable_content>.
- Facebook "Community Standards - Hate Speech" (May 2023) <https://www.facebook.com/communitystandards/hate_speech>.

Facebook Instagram "Community Guidelines" <https://www.facebook.com/help/instagram/477434105621119>.
- Facebook, “Facebook Community Standards” (2024) <https://transparency.meta.com/en-gb/policies/community-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2F >
- Facebook "Community Standards - Oversight Board" (2024) <https://www.facebook.com/communitystandards/oversight_board>.
- Facebook Instagram "Terms of Use" (2025) <https://help.instagram.com/478745558852511?helpref=faq_content>.
- Fiona MacKinnon "Safe Harbour from Insolvency Duties To Expire on 30 September 2020" (2020) <https://kindrik.co.nz/blogs/safe-harbour-from-insolvency-duties-to-expire-on-30-september-2020/>.
- Gary Drenik "Unveiling “X”: The Implications Of Twitter's Bold Rebranding Move" Forbes (8 Septermber 2023) <https://www.forbes.com/sites/garydrenik/2023/09/08/unveiling-x-the-implications-of-twitters-bold-rebranding-move/?sh=455aa9d72ff2>.
- Gilbert Wong “The battle against infodemic threat” (25 October 2022) Mātātaki | The Challenge <https://www.auckland.ac.nz/en/news/2022/10/25/battle-against-infodemic.html>
- Global Internet Forum to Counter Terrorism "Actions to Address the Abuse of Technology to Spread Terrorist and Violent Extremist Content" (2019) <https://gifct.org/press/actions-address-abuse-technology-spread-terrorist-and-violent-extremist-content/>.
- Global Internet Forum to Counter Terrorism "Governance" (2020) <https://gifct.org/governance/>.
- Global Internet Forum to Counter Terrorism "Formation" (2021) <https://gifct.org/about/story/#june-26--2017---formation-of-gifct>.
- Global Internet Forum to Counter Terrorism, “May 2019 Christchurch Call to Action” (2019) <https://gifct.org/about/story/#may-2019---christchurch-call-to-action >
- Global Times "Trump’s racist words spark hatred, fuel global xenophobia " (2020) <https://www.globaltimes.cn/content/1183207.shtml>.
- Greta Yeoman “Too little known about hate-motivated offending - NZLS” (online ed, 2 April 2025) < https://www.capitalletter.co.nz/news/hate-crime/815799/too-little-

known-about-hate-motivated-offending-nzls?utm_source=newsletter&utm_medium=email&utm_campaign=capital-letter-newsletter >

- Hate Aid "Mixed feelings: Digital Services Act replaces NetzDG" (2023) <https://hateaid.org/en/mixed-feelings-digital-services-act-replaces-netzdg/>.
- Human Rights Law Centre “New Zealand High Court finds insulting cartoons did not breach hate speech legislation” (2018) <https://www.hrlc.org.au/human-rights-case-summaries/2018/6/1/new-zealand-high-court-finds insulting-cartoons-did-not-breach-hate-speech-legislation>.
- Ian Wishart "EU Chief Takes Aim at Internet Giants Over Freedom of Speech" (2021) <https://www.bloomberg.com/news/articles/2021-01-26/eu-chief-takes-aim-at-internet-giants-over-freedom-of-speech>.
- Ido Goldberg and others "OpenWeb tests the impact of “nudges” in online discussions" 2020 <https://www.openweb.com/blog/openweb-improves-community-health-with-real-time-feedback-powered-by-jigsaws-perspective-api>
- Jack McKee “Protesters deliver anti-lockdown, vaccine messages to government” (9 November 2021) RNZ <https://www.rnz.co.nz/news/national/455307/protesters-deliver-anti-lockdown-vaccine-messages-to-government>
- James McBride “How does the European Union Work?” (March 2022) Council on Foreign Relations <https://www.cfr.org/backgrounder/how-does-european-union-work>
- Jason Brownlee "What is Deep Learning?" (2020) <https://machinelearningmastery.com/what-is-deep-learning/>.
- Jason Silverstein “The global impact of George Floyd: How Black Lives Matter protests shaped movements around the world” (4 June 2021) < https://www.cbsnews.com/news/george-floyd-black-lives-matter-impact/>
- Jeremy Blackburn and others "Does 'deplatforming' work to curb hate speech and calls for violence? 3 experts in online communications weigh in” (16 January 2021) <https://theconversation.com/does-deplatforming-work-to-curb-hate-speech-and-calls-for-violence-3-experts-in-online-communications-weigh-in-153177>.
- Jeremy Butterfield "Fatwa" in Fowler’s Concise Dictionary of Modern English (2016) <https://www-oxfordreference-com.ezproxy.waikato.ac.nz/view/10.1093/acref/9780199666317.001.0001/acref-9780199666317-e-1367>.

- Jess Berentson-Shaw and Marianne Elliot “Misinformation and Covid-19: a briefing for media” (2020) The Workshop < https://www.theworkshop.org.nz/publications/misinformation-and-covid-19-a-briefing-for-media>
- J. Clement "Countries with the most Twitter users 2020" (2020) <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>.
- Julie de Bailliencourt” Helping creators understand our rules with refreshed Community Guidelines” (Mar 2023) < https://newsroom.tiktok.com/en-us/community-guidelines-update>
- Jordan Valinsky “Elon Musk rebrands Twitter as X” (July 2023) CNN < https://edition.cnn.com/2023/07/24/tech/twitter-rebrands-x-elon-musk-hnk-intl/index>
- John Groom, Natasha Denton and Kathy Harford “European Union: The Digital Services Act - What is changing in the world of tech?” (22 October 2022) Baker McKenzie <https://www.globalcompliancenews.com/2023/10/22/https-insightplus-bakermckenzie-com-bm-technology-media-telecommunications_1-european-union-the-digital-services-act-what-is-changing-in-the-world-of-tech_10172023/>
- Karen Ho “The Facebook whistleblower says its algorithms are dangerous. Here’s why.” MIT Technology Review (5 October 2021) <https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/>
- Karl Schaffarczyk "Explainer: what is geoblocking?" (2021) <http://theconversation.com/explainer-what-is-geoblocking-13057>.
- Kevin R. Davis "John Milton" (6 May 2024) <https://www.mtsu.edu/first-amendment/article/1259/john-milton>.
- Kylie Pappalardo and Nicolas Suzor "Pappalardo, Kylie; Suzor, Nicolas "The Liability of Australian Online Intermediaries" (2018) 40 Sydney Law Review 469.
- Kyriakos Fountoukakos, Susan Black and Kristien Geeurickx “The EU Commission Adopts New Horizontal Cooperation Guidelines And Publishes the R&D And Specialization Block Exemption Regulations Which Introduce A New Era For Horizontal Cooperation In Line With The Court Of Justice’s And The Commission’s Decisional Practices” (1 June 2023) <https://www.concurrences.com/en/bulletin/news-

issues/june-2023/the-eu-commission-adopts-new-horizontal-cooperation-guidelines-and-publishes-113508>.

- Laura Kayali “French constitutional court strikes down most of hate speech law” (2020) <https://www.politico.eu/article/french-constitutional-court-strikes-down-most-of-hate-speech-law/>
- Laurence Helfer and Molly K. Land “Is the Facebook Oversight Board an International Human Rights Tribunal?” (2021) The Lawfare Institute <https://www.lawfaremedia.org/article/facebook-oversight-board-international-human-rights-tribunal>
- Liza Negriff "The Past, Present, and Future of Freedom of Speech and Expression in the People’s Republic of China" (2021) Topical Research Digest: Human Rights In China <https://www.du.edu/korbel/hrhw/researchdigest/china/FreedomSpeechChina.pdf>.
- Madhav Chanchani “India clocks over 5.5 billion hours on TikTok in 2019” (2020) Diligent < https://timesofindia.indiatimes.com/business/india-business/india-clocks-over-5-5-billion-hours-on-tiktok-in-2019/articleshow/73787861.cms>
- Maryam Mohsin "10 Instagram Statistics You Need to Know in 2021 [New Data]" (2021) <https://www.oberlo.com/blog/instagram-stats-every-marketer-should-know>.

- Martina Mantovani "Jurisdiction, Conflict of Laws and Data Protection in Cyberspace" (2017) <https://conflictoflaws.net/2017/jurisdiction-conflict-of-laws-and-data-protection-in-cyberspace/>.
- Markus Reuter and Chris Köver “Cheerfulness and censorship” (2019) <https://netzpolitik.org/2019/cheerfulness-and-censorship/>
- Malik Jitendra Singh and others "Deep Learning for Hate Speech Detection: A Comparative Study" (2023) Ithaca <http://arxiv.org/abs/2202.09517>.
- Mansoor Iqbal “TikTok Revenue and Usage Statistics” (18 April 2024) <https://www.businessofapps.com/data/tik-tok-statistics/>
- Mary Hedengren "Epideictic Rhetoric" (podcast, 5 November 2015) Mere Rhetoric <https://mererhetoric.libsyn.com/epideictic-rhetoric>.
- *Merriam-Webster.com Dictionary* (online ed) < https://www.merriam-webster.com/dictionary/social%20media>.

- Megan Garber "Instagram Was First Called ‘Burbn’" (2 July 2014) The Atlantic <https://www.theatlantic.com/technology/archive/2014/07/instagram-used-to-be-called-brbn/373815/>
- Michael Geist "The Unintended Equustek Effect - How one case set a precedent for Canadian courts' growing jurisdiction over internet activities." (2019) <https://www.cigionline.org/articles/unintended-equustek-effect/>.
- Molly Lesher, Hanna Pawelec and Arpitha Desai, ‘Disentangling untruths online: Creators, spreaders and how to stop them’, (29 March 2022) OECD Going Digital Toolkit Notes <https://goingdigital.oecd.org/en/notes>
- Laura O’Connell Rapira and Leroy Beckett “Our Hate Speech Laws” The People’s Report on Online Hate, Harassment, and Abuse (2018) < https://actionstation.org.nz/publications >.
- M Zain Sarwar “Blaise Pascal’s Invention” (18 February 2024) <https://www.cantorsparadise.com/blaise-pascals-invention-67d21af9cd3b >
- Michael Daubs “Trust, misinformation and social in(ex)clusion” (June 2022) < https://www.royalsociety.org.nz/what-we-do/our-expert-advice/speakers-science-forum/speakers-science-forum-2022/speakers-science-forum-misinformation/>.
- Michelle Duff "Hate crime law review fast-tracked following Christchurch mosque shootings" (2019) Stuff <https://www.stuff.co.nz/national/christchurch-shooting/111661809/hate-crime-law-review-fasttracked-following-christchurch-mosque-shootings>.
- Megan Garber "Instagram Was First Called ‘Burbn’" (2014) <https://www.theatlantic.com/technology/archive/2014/07/instagram-used-to-be-called-brbn/373815/>.
- Meta "Introducing Meta: A Social Technology Company" (2021) <https://about.fb.com/news/2021/10/facebook-company-is-now-meta/>.
- "New Zealand Archives” SocialMedia.org.nz <https://socialmedia.org.nz/category/new-zealand/>.
- Mette Newth "The Long History of Censorship" (2010) < https://brewminate.com/the-long-history-of-censorship/ >
- Natasha Lomas "Understanding Europe's big push to rewrite the digital rulebook" (31 December 2020) <https://techcrunch.com/2020/12/30/understanding-europes-big-push-to-rewrite-the-digital-rulebook/>.

- Nikki MacDonald “Online harassment: the insidious face on an inescapable harm” (11 March, 2019) Stuff <https://www.stuff.co.nz/national/crime/110956646/online-harassment-the-insidious-face-on-an-inescapable-harm>.
- Nikki Mizuguchi “VPN growth highlights global crackdown on internet freedom” (2023) < https://asia.nikkei.com/Business/Technology/VPN-growth-highlights-global-crackdown-on-internet-freedom >
- Nicola Gray “Social Media and The Use of ‘Clout’” (January 14, 2020) <http://www.theeditgcu.com/arts-culture/social-media-and-the-use-of-clout/>.
- Neil Melhuish and Edgar Pacheco “*Measuring trends in online hate speech victimisation and exposure, and attitudes in New Zealand*” (December 12, 2019) Social Science Research Network , <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3501977> at 1.
- Netsafe "What is the HDCA?" (21 September 2021) Bullying & Abuse; Online Safety Parent Toolkit < https://netsafe.org.nz/what-is-the-hdca/>
- New Zealand Immigration Law "Relationship support letter guide" (2019) <https://nzil.co.nz/relationship-support-letter-guide/>.
- New Zealand Parliament "Films, Videos and Publications Classification (Urgent...Harm) Amendment Bill - Digest 2626 - New Zealand Parliament" (2021) <https://www.parliament.nz/en/pb/bills-and-laws/bills-digests/document/52PLLaw262611/films-videos-and-publications-classification-urgentharm>.
- Nick Clegg "Facebook Does Not Benefit from Hate - About Facebook" (2020) <https://about.fb.com/news/2020/07/facebook-does-not-benefit-from-hate/>.
- Oliver Bell and Vito Petretti "European Commission Publishes the Digital Services Act" (29 December 2020) Lexology Tech & Sourcing @ Morgan Lewis <https://www.lexology.com/library/detail.aspx?g=913db9d9-b813-4520-b827-97365669b8f0>
- “Online harassment: the insidious face on an inescapable harm” <https://www.stuff.co.nz/national/crime/110956646/online-harassment-the-insidious-face-on-an-inescapable-harm>.
- Oversight Board "Oversight Board demands more transparency from Facebook" (2021) <https://www.oversightboard.com/news/215139350722703-oversight-board-demands-more-transparency-from-facebook/>.

- PBS News “Supreme Court upholds TikTok ban if not sold by Chinese, Trump has promised a solution” (17 January 2025) PBS News < https://www.pbs.org/newshour/politics/supreme-court-upholds-tiktok-ban-if-not-sold-by-chinese-trump-has-promised-a-solution>
- Pete Mitchell "Can a Deleted FB Account be Traced?" (14 June 2023) <https://techcult.com/can-a-deleted-fb-account-be-tr>.
- Peter Warren Singer and Emerson T Brooking “LikeWar: The weaponization of social media” Eamon Dolan Books (2018) <https://www.proquest.com/trade-journals/likewar-weaponization-social-media/docview/2124045248/se-2>.
- Philip Oltermann "Obscure German law gives Angela Merkel a diplomatic headache" (2016) <http://www.theguardian.com/world/2016/apr/14/obscure-german-law-angela-merkel-recep-tayyip-erdogan>.
- Robert V Labaree “Research Guides: Organizing Your Social Sciences Research Paper: Theoretical Framework” <https://libguides.usc.edu/writingguide/theoreticalframework>.
- Rachel Ranosa "How recruiters check for red flags on social media" (2019) <https://www.hcamag.com/nz/specialisation/hr-technology/how-recruiters-check-for-red-flags-on-social-media/189900>.
- Rachel Sadler “In-fighting between Freedom and Rights Coalition, Counterspin continues at convoy protest after event ‘hijacked’” (2022) Newshub < https://www.newshub.co.nz/home/new-zealand/2022/02/in-fighting-between-freedom-and-rights-coalition-counterspin-continues-at-convoy-protest-after-event-hijacked.html>.
- Rawiri Tuapawa (@RaTheKenDoll) <www.tiktok.com/rathekendoll>
- Richard Allan "Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community? - About Facebook" (2017) <https://about.fb.com/news/2017/06/hard-questions-hate-speech/>.
- Radio New Zealand (RNZ) 1995 (NZ) <https://www.rnz/about.>
- RNZ "Little plans fast-track review of hate speech laws" (2019) <https://www.rnz.co.nz/news/national/385955/little-plans-fast-track-review-of-hate-speech-laws>.
- Sarah Perez "TikTok takes down some hashtags related to election misinformation, ignores others" (2020) TechCrunch < https://techcrunch.com/2020/11/05/tiktok-takes-down-some-hashtags-related-to-election-misinformation-leaves-others/>.

- Sarah Perez "TikTok expands Community Guidelines, rolls out new ‘well-being’ features -" (2020) TechCrunch < https://techcrunch.com/2020/12/15/tiktok-expands-community-guidelines-rolls-out-new-well-being-focused-features/>.
- Sarah Perez "Twitter expands hateful conduct rules to ban dehumanizing speech around age, disability and now, disease - TechCrunch" (2020) <https://social.techcrunch.com/2020/03/05/twitter-bans-hate-speech-around-age-disability-and-in-the-wake-of-the-coronavirus-outbreak-disease/>.
- Sam Biddle, Paulo Victor Ribeiro and Tatiana Dias “Invisible Censorship - TikTok Told Moderators to Suppress Posts by “Ugly” People and the Poor to Attract New Users” (2020) The Intercept <https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/>
- Samantha Cole “Where Did the Concept of 'Shadow Banning' Come From?” (2018) <https://www.vice.com/en/article/a3q744/where-did-shadow-banning-come-from-trump-republicans-shadowbanned>.
- Simon Kemp “Social media users pass the 4 billion mark as global adoption soars” <https://wearesocial.com/cn/blog/2020/10/social-media-users-pass-the-4-billion-mark-as-global-adoption-soars/>
- Stephen Warwick "Facebook is changing its Terms of Service, and users are not happy" <https://www.windowscentral.com/facebook-changing-its-terms-service-and-users-are-not-happy>.
- Syed Aftab Hassan Bukhari “What is Comparative Study” (November 20, 2011). Social Science Research Network <http://dx.doi.org/10.2139/ssrn.1962328>
- Stacy Jo Dixon "Number of social media users worldwide 2010-2021" (April 2020) Statista <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>.
- Shanti Mathias, ‘Tracking the surge in online anti-trans hate sparked by Posie Parker’s visit’, date? (5 May 2023) The Spinoff < https://thespinoff.co.nz/internet/05-05-2023/tracking-the-surge-in-online-anti-trans-hate-sparked-by-posie-parkers-visit>.
- Stephen Warwick "Facebook is changing its Terms of Service, and users are not happy" (1 September 2020) Windows Central <https://www.windowscentral.com/facebook-changing-its-terms-service-and-users-are-not-happy>
- Sofia Cherici “Mirroring Bias: Online Hate Speech and Polarisation” *Green European Journal* 2021 < https://www.greeneuropeanjournal.eu/mirroring-bias-online-hate-speech-and-polarisation/>

- Skylar Hughes “Lateral reading: The best media literacy tip to vet credible sources” (20 July 2023) Poynter. < https://www.poynter.org/fact-checking/media-literacy/2023/lateral-reading-the-best-media-literacy-tip-to-vet-credible-sources/>.
- Sharon Brettkelly “Parker’s visit poses plenty of questions” (30 March 2023) RNZ <https://www.rnz.co.nz/programmes/the-detail/story/2018883814/parker-s-visit-poses-plenty-of-questions>.
- Tom Hunt “By the numbers: The 23 days of New Zealand’s Parliament occupation” (2023) Stuff <https://www.stuff.co.nz/dominion-post/news/wellington/131356257/by-the-numbers-the-23-days-of-new-zealands-parliament-occupation>.
- TikTok "Community Guidelines" (2020) <https://www.tiktok.com/community-guidelines?lang=en>.
- TikTok "What is the 'For You' feed?" (2023) <https://www.tiktok.com/creators/creator-portal/en-us/how-tiktok-works/whats-the-for-you-page-and-how-do-i-get-there/>.
- TikTok “Countering Hate on TikTok” (2023) <https://www.tiktok.com/safety/en/countering-hate/>
- The Associated Press "Landmark European Court Ruling Could Force Google, Yahoo, Bing To Scrub Online Reputations" (2014) <https://www.cbsnews.com/sanfrancisco/news/landmark-european-court-ruling-could-force-google-yahoo-bing-to-scrub-online-reputations/>.
- The Tor Project "The Tor Project" <https://www.torproject.org/about/history/>.
- Te Aka Māori Dictionary (2024) <https://maoridictionary.co.nz/search?idiom=&phrase=&proverb=&loan=&histLoanWords=&keywords=takatapui>.
- TikTok “What is TikTok LIVE?” (2020) <https://support.tiktok.com/en/live-gifts-wallet/tiktok-live/what-is-tiktok-live>.
- Twitter "Hateful Conduct Policy" (2021) <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>.
- Twitter "Twitter Community Guidelines" (April 2023) <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>.
- Twitter "Our range of enforcement options" (2021) <https://help.twitter.com/en/rules-and-policies/enforcement-options>.
- TikTok "Community Guidelines" (March 2023) <https://www.tiktok.com/community-guidelines?lang=en>.
- Instagram <https://www.instagram.com/>.

- Twitter <https://twitter.com/>.
- TikTok <https://www.tiktok.com/en/>.
- Thomas de Weerd and Jurre Reus "News Update - The Digital Services Act And The Digital Markets Act" (2021) <https://www.houthoff.com/insights/News-Update/News-Update-The-Digital-Services-Act-and-the-Digital-Markets-Act>.
- Waatea news.com "Deputy Prime Minister role to be shared" (2023) <https://waateanews.com/2023/11/26/deputy-prime-minister-role-to-be-shared/>.
- WARC, "TikTok India forecasts 'at least' 50% user growth in 2020" (2020) < https://www.warc.com/newsandopinion/news/tiktok-india-forecasts-at-least-50-user-growth-in-2020/en-gb/43327>
- William Perrin and Maeve Walsh Lorna Woods "Draft Online Harm Reduction Bill" (2019) <https://www.carnegieuktrust.org.uk/publications/draft-online-harm-bill/>.
- William Perrin "Government online harms proposals reflect Carnegie UK Trust work" (2021) Government online harms proposals reflect Carnegie UK Trust work <https://www.linkedin.com/pulse/government-online-harms-proposals-reflect-carnegie-uk-william-perrin?trk=public_profile_article_view>.
- Yuxi Wang and others "Understanding and neutralising covid-19 misinformation and disinformation" (2022) BMJ 379 <https://www.bmj.com/content/bmj/379/bmj-2022-070331.full.pdf>
- 1News "What are Posie Parker's views and why are they so controversial?" (24 March 2023) 1News < https://www.1news.co.nz/2023/03/24/what-are-posie-parkers-views-and-why-are-they-so-controversial/>

Other resources

- Abby Vesoulis How Gab Became the Social Media Site Where the Pittsburgh Suspect's Anti-Semitism Thrived (Time Magazine, (2018).
- Andrew Chen "Submission on the Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2021"
- Amanda Meade "Australia is making Google and Facebook pay for news: what difference will the code make?" *The Guardian* (9 Dec 2020)

- Claire Mason and Kathy Errington, ‘Claire Mason and Katherine Errington "Anti-social media: reducing the spread of harm content on social media networks" (2021) https://helenclark.foundation/publications-and-media/anti-social-media/
- Dan Jerker B. Svantesson "Are we stuck in an Era of Jurisdictional Hyper-regulation" *50 Years of Law and IT* (Stockholm Institute for Scandinavian Law, 2018)
- Eddie L. Ungless, Nina Markl, and Björn Ross, “Experiences of Censorship on TikTok Across Marginalised Identities” (2024) arXiv:2407.14164
- Ella Duggan and Raya Hotter “Covid-19 Omicron outbreak, Parliament protest: Wellington students kept away from school” *New Zealand Herald* (online ed, 2022) <https://www.nzherald.co.nz/nz/covid-19-omicron-outbreak-parliament-protest-wellington-students-kept-away-from-school/XIWT3FQTQJ6BI42Y3EGAOQMLFY/>
- Farhan Asif Chowdhury, Dheeman Saha, Md Rashidul Hasan, Koustuv Saha , Abdullah Mueen "Examining Factors Associated with Twitter Account Suspension Following the 2020 U.S. Presidential Election." (2021) Arxiv <https://doi.org/10.48550/arXiv.2101.09575>
- Federal Ministry of Family Affairs and Federal Ministry of Justice and Consumer Protection (2017) <https://www.bmjv.de/SharedDocs/Pressemitteilungen/DE/2017/03142017_Monitoring_SozialeNetzwerke.html>
- Ian Bassett "Is hate speech legislation necessary or desirable?"*DayStar Magazine* (Online ed, Auckland, 2005)
- Jialun 'Aaron' Jiang and others "Characterizing Community Guidelines on Social Media Platforms" (paper presented to Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing, Virtual Event, USA, 2020).

- Maxim Institute Podcast "4. David Hall and Alex Penk on hate speech and free speech - Maxim Institute Podcast - Podcast" Podtail <https://soundcloud.com/user-864022290/4-david-hall-and-alex-penk-on-hate-speech-and-free-speech>.
- Michael Legg and Felicity Bell *Artificial Intelligence and The Legal Profession: A Primer* (University of New South Wales, Sydney, 2017). <https://allenshub.unsw.edu.au/news/artificial-intelligence-and-legal-profession-primer>

- New Zealand Council for Civil Liberties “Submission on the Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2021"
- Paul Rishworth “Interpreting and Invalidating Enactments Under a Bill of Rights” in Rick Bigwood (ed) The Statute: Making and Meaning (LexisNexis, Wellington, 2004) 251
- Peter Coe "The Draft Online Safety Bill and the regulation of online harms and hate speech: have we opened Pandora’s Box?" (paper presented to BACL Annual Webinar: The Regulation of Hate Speech Online and Its Enforcement in a Comparative Perspective, London, 31 August 2021).
- Rachel Tan “Submission on the Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill 2021".
- Robin Barendze “The (In)visibility of Hobson’s Pledge: A Struggle for Survival in the Sociopolitical Environment of Aotearoa/New Zealand” (Master of Arts in Sociology, Massey University, New Zealand, 2018)
- Royal Commission of Inquiry Into The Terrorist Attack On Christchurch Mosques On 15 March 2019
- Sam Alexander, “A Uniquely Australian Approach: A Thematic Analysis of the Normative Foundations of Australia’s Approach to the Regulation of the Internet” (2022) 43(1) Adelaide Law Review 345
- Sanjana Hattotuwa and Kayli Taylor Kate Hannah "Working Paper, Mis- and disinformation in Aotearoa New Zealand" (The Disinformation Project, 2021)
- Livestream TikTok @RaTheKenDoll (May 2021) <www.tiktok.com/rathekendoll>
- Katharine Gelber "A better way to regulate online hate speech: require social media companies to bear a duty of care to users" (14 July 2021) <https://theconversation.com/a-better-way-to-regulate-online-hate-speech-require-social-media-companies-to-bear-a-duty-of-care-to-users-163808>.
- Katherine Herrick “Breaking Things: Origins and Consequences of Racialized Hate Speech on Facebook” (International Studies Honors Projects, Macalester College, 2022)
- Khavin Dmitriy and others "U.S. Capitol Riot" *The New York Times* (19 January 2021)
- Willaim M. Reilly “Roundup: Guterres warns against COVID-19 misinformation as UN leverages supplies to Africa” Xinhua Net (online ed, China, 15 April 2020).

- Peter Billie Larsen and Marjorie Pamintuan “The Human Right to Science: From Fragmentation to Comprehensive Implementation?” (Research Paper, No. 163, South Centre, Geneva 19 August 2022)