Social Media Platforms Duty of Care – Regulating Online Hate Speech*

Rachel Tan

PhD Candidate (University of Waikato, New Zealand)

* Double-blind reviewed article.

Abstract There has been a proliferation of social media usage over the past decade. Social media platforms offer a convenient mode for virtual social interaction by providing relatively simple access to most people. However, there has been a recurring theme of harassment or bullying by way of hateful speech that causes social harm especially following the Christchurch terrorist attacks in 2019. New Zealand is at a point of inflexion when it comes to updating its laws to combat online hate speech. The manner in which statutory duty of care is proposed as law in other comparative jurisdictions (such as the UK and Australia) will be explored in order to establish whether it is beneficial and necessary to be adopted in New Zealand.

INTRODUCTION

Social media platforms played a paramount role during the Christchurch terrorist events in March 2019. Facebook and Twitter came under public scrutiny on whether it had done enough to stop all harm that arose from its livestreams.[1] New Zealand is at a

[1] Jenni Marsh and Tara Mulholland, 'How the Christchurch terrorist attack was made for social media'. *CNN Business*, 15 March 2019. Accessed at: <https://www.cnn.com/2019/03/15/tech/christchurch-internet-radicalization-intl/index.html>.

point of inflexion when it comes to updating its laws to combat online hate speech.[2] The scope of this Article is specifically on the concept of duty of care and whether a statutory one ought to be imposed onto social media platforms.

The manner statutory duty of care is proposed as law for the regulation of online hate speech in online platforms will be explored in two other comparative jurisdictions, United Kingdom and Australia. These two jurisdictions possess a comprehensive review of the relevant issues, best practices, and literature.

Considering that the landscape of social media has changed over the past decade, this too has changed the way people communicate. Especially in a time of a pandemic, social media has been used to communicate and as a form of escapism.[3] While social media giants such as Facebook and Twitter offer a convenient mode for virtual social interaction by providing relatively simple access to most people, there has been a recurring theme of harassment or bullying by way of hateful speech that causes social harm[4]. In early 2020, the Covid-19 coronavirus pandemic also resulted in a rise of online hate sentiments directed at people of migrant background and of Chinese ethnicity[5]. Many people of Asian background have come forward to indicate that there is presence on social media platforms of Anti-Chinese sentiment which disparages Chinese people[6]. This created a space for social media to step-up and be held accountable.

With all the negativity that social media has caused, it is therefore crucial to examine the existing legal framework and establish if accountability (whether it lies on the end-

[2] Jacinda Ardern, New Zealand Governement, The Beehive Press Relese, 'Significant progress made on eliminating terrorist content online', 24 September 2019. Accessed at: <http://www.beehive.govt.nz/release/significant-progress-made-eliminating-terrorist-content-online>.

[3] Rachel Sue Yin Tan, 'Disabling access to illegal online content by way of takedowns'. *New Zealand Law Journal,* 10 2021, pp.341.

[4] Nikki Macdonald, 'Online harassment: the insidious face on an inescapable harm'. *Stuff*, 11 March 2019. Accessed at: <https://www.stuff.co.nz/national/crime/110956646/online-harassment-the-insidious-face-on-an-inescapable-harm>.

[5] Global Times, 'Trump's racist words spark hatred, fuel global xenophobia'. *Global Times*, 20 March 2020. Accessed at: < https://www.globaltimes.cn/content/1183207.shtml>.

[6] New Zealand Human Rights Commission, 'Meng Foon: Covid-19 coronavirus fear no excuse for racism'. Accessed at: <https://www.hrc.co.nz/news/meng-foon-covid-19-coronavirus-fear-no-excuse-racism>.

user or social media platforms) are set to curtail online hate.[7] Irrespective of the strategies social media companies are attempting to deploy, it does not seem to fix the situation.

To obtain a greater chance for success for the regulation of online hate speech, synergic regulation is key. Lessig's regulation theory indicates that regulating cyberspace is not only a legal problem, but it is also problem to end-users because coded software can affect and regulate the way people behave.[8] Murray further elaborates that by virtue of a dynamic regulatory model, regulators can design a synergic regulation with the pre-existing software infrastructure thereby creating a greater likelihood for success.[9]

In the Christchurch shootings, social media was used in the planning and aftermath of the events to distribute and disseminate images of the attacks. It was at the Global Internet Forum to Counter Terrorism ('GIFCT') that Prime Minister Jacinda Ardern and President of the French Republic Emmanuel Macron announced the implementation of the Christchurch Call to Action ('the Call') at the United Nations General Assembly.[10] The Call was adopted by Heads of States along with technology sector companies.[11] It was also announced that given the existing objectives to 'share knowledge and support research on terrorists' use of platforms'[12] the GIFCT will be relaunched and will become an independent body with new commitments set forth in the 'nine-point action plan'.[13]

[7] Mathew Binny, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal and Animesh Mukherje 'Thou Shalt Not Hate: Countering Online Hate Speech', *Proceedings of the International AAAI Conference on Web and Social Media*, (2019) 13(1).

[8] Lawrence Lessig, 'The New Chicago School'. *The Journal of Legal Studies*, 27(2), 1998, pp. 661-691.

[9] Andrew Murray, *The Regulation of Cyberspace*. London: Routledge-Cavendish, 2007.

[10] Ardern, *Significant progress*.

[11] Edgar Pacheco and Neil Melhuish '2019 online hate speech insights', Netsafe – Online Safety Help and Advice for New Zealanders. Accessed at: <https://www.netsafe.org.nz/2019-online-hate-speech-insights/>.

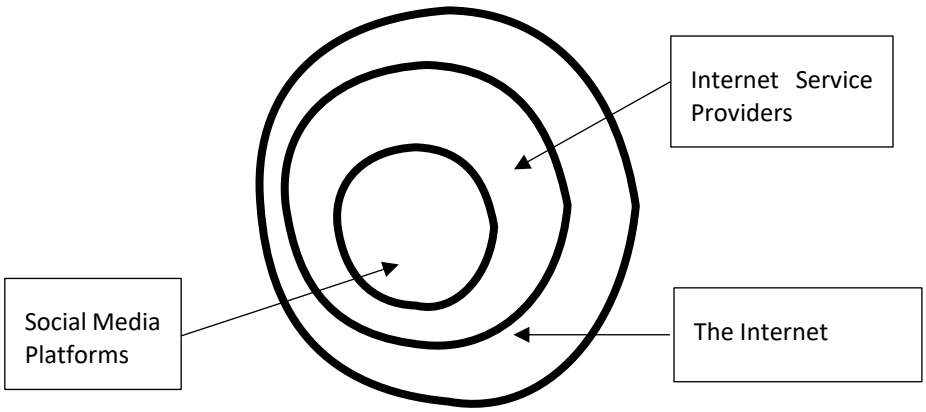
[12] Ardern, *Significant progress*.

[13] Global Internet Forum to Counter Terrorism, 'Actions to Address the Abuse of Technology to Spread Terrorist and Violent Extremist Content'. Accessed at: <https://gifct.org/press/actions-address-abuse-technology-spread-terrorist-and-violent-extremist-content/>.

HOW INTERMEDIARIES REGULATE ONLINE HATE SPEECH

An Internet intermediary is an entity which provides services that enable people to use the internet.[14] These are of two classes, conduits – which are technical providers of internet and hosts – which are providers of content.[15] Internet Service Providers are examples of conduit intermediaries, while Facebook and Twitter would be examples of hosts intermediaries. Internet intermediaries are technically designed to permit storage, creation of content and transmission of information.[16]

Figure 1. The relationship between social media platforms and Internet Service Providers.



Social media is given an atmosphere to function within an Internet Service Provider ('ISP') as shown in the diagram above. Considering that online hate speech exists in social media platforms, we should examine if liability should exist for ISPs as well.

[14] Association for Progressive Communications, 'Frequently asked questions on internet intermediary liability' Association for Progressive Communications. Accessed at: <https://www.apc.org/en/pubs/apc%E2%80%99s-frequently-asked-questions-internet-intermed>.

[15] Association for Progressive Communications, *Frequently asked questions*.

[16] Jaani Riordan, *The Liability of Internet Intermediaries.* Oxford: Oxford University Press, 2016.

Default mechanisms such as censorships, geo-blocking, web-filters and takedown of hateful content are used to help curtail online hate speech.[17] Community Guidelines have also been developed for this purpose.[18]

Community Guidelines have become a reference point for the way users behave and conduct themselves in respective social media spaces. They comprise of a set of rules laid out by respective social media platforms which enforce governance as a passive approach to moderating content.[19] This means that if a user acts in a manner that contravenes the Community Guidelines or rules, there will be a consequence. Examples of offences that can contravene community guidelines are cyberstalking, misusing intellectual property and of course, objectionable content in which online hate speech falls under. It is important to have community guidelines in place to ensure that the social media environment is a safe place for its users to interact and express themselves.

All of these mechanisms have been put into place by host intermediaries in an effort to self-regulate. However, these intermediaries were not being held accountable by existing laws. With the prolific expansion of the internet, which occurred during the late 2000s, national and international institutions expanded its regulations thus creating new liability rules.[20] The expansion developed new forms of secondary liability. As online content grew, there were also enforcement problems. This brought a dire need for stronger enforcement bringing new limitations to the fundamental rights of intermediaries and its users.[21]

However, there is a question on whether conduit intermediaries should also share accountability. From its early days, ISPs have resisted to be stifled by a legislative framework that would hold them accountable and liable.[22] The rationale and argument

[17] Rachel Sue Yin Tan, 'Disabling access to illegal online content by way of takedowns'. *New Zealand Law Journal,* 10, 2021, pp.341.

[18] Barbara Perry and Patrik Olssen, 'Cyberhate: The globalization of hate'. *Information & Communications Technology Law*, 18(2), 2009, pp. 185-199.

[19] Jialun 'Aaron' Jiang, Skyler Middler, Jed R. Brubaker and Case Fiesler, 'Characterizing Community Guidelines on Social Media Platforms' *Association for Computing Machinery Digital Library.* Accessed at: <https://doi.org/10.1145/3406865.3418312>.

[20] Riordan, *Liability of Intenet*, p. 15.

[21] Riordan, *Liability of Intenet*, p. 15.

[22] E. Eugene Clark, *Cyber law in Australia*, Kluwer Law International, 2010, p.318.

for not having a legislative code for ISPs was that they viewed themselves as bookshops, libraries, and postal workers – in that they obviously would not have any knowledge of the contents of its entire catalogue of books, or its contents in envelopes.[23] In principle, an ISP may be primarily liable when it has knowledge, control or financial benefit for the information or content:[24] knowledge being the key factor.

Internet intermediaries can now be identified as 'Authority Gatekeepers'.[25] An internet service intermediary encompasses a relationship between infrastructure providers, the platform, small intermediaries (such as an Administrator of a Facebook Page) and the receptor or end-users (who could also be a creator of speech).[26]

With the evolution of the internet, internet intermediaries have become a vital and dependable part of any critical national infrastructure such as healthcare, communications, finance, food, public services, energy, and transportation.[27] Therefore, it has been in the best interest of governments to create regulatory frameworks to protect governmental institutions, businesses, and the general public from harm. This has changed the liabilities of internet intermediaries within the legal framework.

DUTY OF CARE IN THE SOCIAL MEDIA SPHERE

The concept of a legal 'duty of care' has been influenced by common law, and more recently, codified by statute in New Zealand. Questions relating to which entities owe a legal duty of care to which end-users, as well as the nature of that duty, are complex – particularly when it comes to social media platforms, which can be simultaneously described as 'intermediaries', 'services' and 'products', depending on the context.

[23] Riordan, *Liability of Intenet*, p. 38.

[24] Clark, *Cyberlaw in Australia*, p. 314.

[25] Emily Laidlaw, 'Internet Gatekeepers, Human Rights and Corporate Social Responsibilities'. *London School of Economics and Political Science*. Accessed at: <http://etheses.lse.ac.uk/317/>.

[26] Laidlaw, *Internet Gatekeepers*, p. 317.

[27] United Kingdom Government, Cabinet Office, 'Cyber Security Strategy of the United Kingdom: Safety, Security and Resilience in Cyber Space'. Accessed at: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/228841/7642.pdf>.

One common approach is to conceptualise social media platforms as a form of internet intermediaries (a conduit between two or more individuals interacting with each other). Hutty, the Chair of EuroISPA's Intermediary Liability Committee, has a simple approach to describing the notion of limited liability of intermediaries based on a distinction between a *service* (such as a telecommunications service correctly described as an intermediary) and a *publisher* (such as a Newpaper Outlet that is not an intermediary).[28] According to this concept, when it comes to harm caused to an end-user by the action of another actor, intermediaries should be protected from liability. However, if the social media platform fails to meet the definition of an 'intermediary' – because, for example, it becomes seen as more actively involved in generating and distributing content - it is stripped of the protection that prevents them from being treated as though they are publishers.[29] This in turn has implications for the legal duties owed by the platform to its end users, including with respect to providing protection from online hate speech.

Whether or not any particular social media platform will be treated as an intermediary or publisher depends on the jurisdiction, and different national standards within and across jurisdictions.[30] This gives rise to significant complexity for end-users around the world, seeking to understand their legal rights when it comes to remedies for harm caused by online hate speech.

The European Commission President, von der Leyen, said that there ought to be a single legal framework that would stipulate the responsibility for the manner internet intermediaries:

> *disseminate, promote, and remove content…(sic). We want the platforms to be transparent about how their algorithms work*

[28] EuroISPA is the world's largest association of internet service providers. See EuroISPA, 'Recap of Past Event: Liability of Intermediaries'. Accessed at: <https://www.euroispa.org/2021/10/recap-of-past-event-liability-of-intermediaries/>.

[29] EuroISPA, *Liability of Intermediaries*.

[30] EuroISPA, *Liability of Intermediaries*.

> *because we cannot accept that decisions that have a far-reaching impact on our democracy are taken by computer programs alone.*[31]

This emphasises the European Union's position that social media platforms should take on more accountability and embrace a duty of care approach, that more accurately recognises their role as providing both a service and a product to end users.[32]

The landmark case from tortious law, *Donoghue v Stevenson*[33] provides an example of how the common law approach to 'duty of care' could be applied in the context of social media platforms and online hate speech. In this case, the claimant drank a bottle of ginger beer that was purchased by her friend at a café.[34] Upon finishing her beverage, the claimant found a decomposing snail inside the bottle. She had not noticed the snail in the bottle beforehand as the bottle was opaque, and as a result, she fell ill and suffered nervous shock and gastroenteritis. In this case, the producer of the ginger beer was the defendant, Stevenson. Among the several issues arising in the case, were the following three questions:

- Whether there was a legal duty of care owe by Stevenson as producer of the ginger beer to Donoghue as the consumer.
- Whether it was relevant that Donoghue had not purchased the ginger beer and that her friend was the actual purchaser.
- Whether Donoghue had locus standi to bring the claim against Stevenson

These questions – relating to the scope of duty of care owed to purchasers and consumers - also arise in the context of users interacting with social media platforms, particularly if social media platforms are seen as offering a 'product' rather than merely being an 'intermediary' or forming part of a service. In this way, the findings made in

[31] Ian Wishart, 'EU Chief Takes Aim t Internet Giants Over Freedom of Speech'. *Bloomberg News*, 26 January 2021. Accessed at: <https://www.bloomberg.com/news/articles/2021-01-26/eu-chief-takes-aim-at-internet-giants-over-freedom-of-speech>.

[32] Ian Wishart, *Internet Giants*. 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.

[33] *Donoghue v Stevenson* [1932] UKHL 100 '*Donoghue v Stephenson*'.

[34] *Donoghue v Stevenson*.

Donoghue v Stevenson can be drawn upon to conceptualise the legal responsibility owed by social media companies to the users of their platforms.[35]

For example, in the case of *Donoghue and Stevenson*, the Lord Atkin held that 'a manufacturer of products, which he sells…to reach the ultimate consumer in the form in which they left him…, owes a duty to the consumer to take reasonable care'.[36] In another landmark torts case, *Bourhill v Young*, Lord Thankerton observed that:

> *The English cases demonstrate how impossible it is to catalogue finally, amid the ever-varying types of human relationships, those relationships in which a duty to exercise care arises apart from contract, and each of these cases relates to its own set of circumstances, out of which it was claimed that the duty had arisen. In none of these cases were the circumstances identical with the present case as regards that which I regard as the essential element in this case, namely, the manufacturer's own action in bringing himself into direct relationship with the party injured. I have had the privilege of considering the discussion of these authorities by my noble and learned friend Lord Atkin in the judgment which he has just delivered, and I so entirely agree with it that I cannot usefully add anything to it.*[37]

This suggest that, in the context of social media platforms such as Facebook, Twitter and TikTok, a duty of care extends to the end-user of a social media 'product' and that when discharging that duty, reasonable care must be taken to protect end users from harm, including harm caused by online hate speech.

If a common law duty of care does exist between social media platforms and their users, which extends to a duty to take reasonable care to protect users from online hate speech, this could play an important role in addressing some of the short comings

[35] Kylie Pappalardo and Nicolas Suzor, 'The Liability of Australian Online Intermediaries'. *Sydney Law Review*, 40(4) 2018, pp.469.

[36] *Donoghue v Stevenson*.

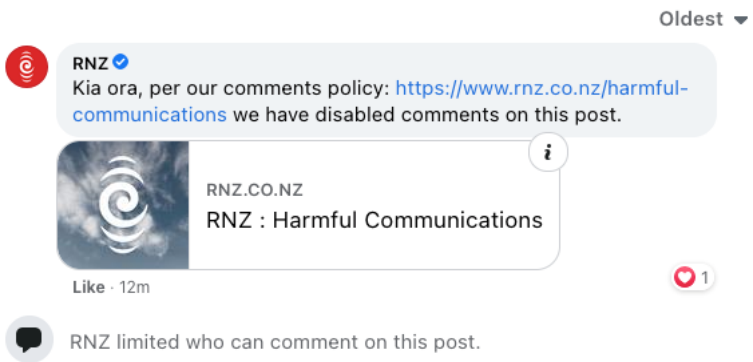
[37] *Bourhill v Young* [1943] AC 92 at 603. ('*Bourhill v Young*').

arising from the largely 'self-regulated' approach to protecting social media users currently evident in New Zealand and Australia.

Self-regulation by content hosts has become prevalent on social media platforms as demonstrated in the image below. It depicts Radio New Zealand (RNZ),[38] being a host on Facebook, self-regulating its comment section in line with its obligations under the *Harmful Digital Communications Act 2015* (NZ) ('HDCA').[39] The intention and purpose of the HDCA is to protect users from harm caused over the internet, but the legislation relies on a predominantly 'self-regulated' approach to enforcement of and compliance with safety standards by social media platforms and content hosts.

In the below example, RNZ sought to implement its responsibilities under the HDCA by adding a 'formal and visible warning on all our platforms so the public is aware of what…'[40] will occur should the lines get crossed. In addition, as a content host, RNZ initiated switching comments off on posts that had a likelihood of either abusive or harmful comments. RNZ also refers to Facebook's Community Guidelines when taking these actions, taking an active role in self-regulating its content on Facebook.[41]

Figure 2. Diagram 2: RNZ's Comment Section on Facebook



[38] 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>.

[39] *Harmful Digital Communications Act 2015* (NZ).

[40] Radio New Zealand, '*Harmful Communications*'. Accessed at: <https://www.rnz.co.nz/harmful-communications>.

[41] Radio New Zealand, *Harmful Communications*.

As a host RNZ urges its users to contemplate the following questions prior to publishing on any of its platforms: 'Ask yourself: would this offend someone? Is it defamatory? How would you react if someone else wrote the same thing?'.[42] This could be seen as an attempt by RNZ to discharge its responsibilities under the HDCA, or alternatively, as a way of shifting the 'duty of care' from the host to the user. Either way, this example demonstrates the clear limitations on the effectiveness of self-regulation as a form of protection from online hate speech and highlights the need to consider imposing enforceable statutory obligations on key actors within the social media sphere.

A STATUTORY DUTY OF CARE ONTO SOCIAL MEDIA PLATFORMS - UNITED KINGDOM

The United Kingdom has experimented with imposing statutory duties of care on social media platforms, with mixed success. In 2019 the UK Parliament considered the Online Harm Reduction Bill which proposed a comprehensive new legal framework imposing a new statutory duty of care on social media platforms.[43] The Bill is currently on the Report Stage in Parliament as of 12th of July 2022.[44]

The Online Harm Reduction Bill features a set of safety standards and statutory duties influenced by the Health and Safety at Work Act.[45] The Bill also proposes the formation of a separate and independent body to enforce those duties, the Office of Communications ('OFCOM'),[46] which is also tasked with developing codes practice in consultation with key industry stakeholders. The proposed OFCOM aims to provide a regulatory body that can take steps to reduce this harm by enforcing a statutory duty of care that is owed to every user of online platforms, including Facebook, Instagram, Twitter and TikTok. The Bill takes a 'deliberately consultative and iterative approach in

[42] Radio New Zealand, *Harmful Communications*.

[43] Lorna Woods, 'The duty of care in the Online Harms White Paper'. *Journal of Media Law*, 11(1), 2019, pp. 6-17.

[44] UK Parliament, 'Parliamentary Bills – Online Safety Bills'. Accessed at: <https://bills.parliament.uk/bills/3137/stages/16798>

[45] Lorna Woods, William Perrin and Maeve Walsh, 'Draft Online Harm Reduction Bill: Explanatory Memorandum', Carnegieu UK Trust. Accessed at: <https://www.carnegieuktrust.org.uk/publications/draft-online-harm-bill/>.

[46] House of Lords, Select Committee on Communications, Parliament of United Kingdom, *Regulating In A Digital World*, 2nd Report, Session 2017-19.

developing the framework'.[47] This was designed to ensure that the legislation is 'coherent, proportionate and agile in response to advances in technology'.[48]

The Online Harm Reduction Bill imposes a legal duty on social media operators equivalent to the duty imposed on an employer under the Health and Safety at Work Act, based on the 'safety by design' approach.[49] It is designed to ensure social media platforms provide safe and healthy conditions and protect their users from stress or bullying in the design of the platform.

One of the benefits of the Online Harm Reduction Bill model is that it clearly recognises that hate speech in social media is in fact a form of harm. The Bill also recognises that there is a wide spectrum of harm that can occur within the online environment and there may be a need for specific harm reduction mechanisms that protect vulnerable groups.[50] Imposing a duty of care on larger corporations such as Facebook and Instagram aims to compel these entities to identify the *harm* by way of taxonomy and to take reasonable steps to mitigate such *harms.*[51] On the other hand, while this proposed approach may provide a framework to better understand the nature of the harm caused, it has not yet delivered a clear set of prescriptive rules or standards to follow.[52] This has compelled some social media platforms to continue to ascertain and improve their responses to harm reduction within a regulatory framework better suited to other forms of nationally-controlled telecommunication services albeit voluntarily.[53]

[47] Secretary of State of Digital, Culture, Media and Sport, *Consultation Outcome - Online Harms White Paper: Full government response to the consultation,* December 2020*.* Accessed at: <https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response>.

[48] Secretary of State of Digital, Culture, Media and Sport, *Consultation Outcome*.

[49] Secretary of State of Digital, Culture, Media and Sport, *Consultation Outcome*.

[50] William Perrin, 'Government online harms proposals reflect Carnegie UK Trust work', Linked In post, 5 January 2021. Accessed at: <https://www.linkedin.com/pulse/government-online-harms-proposals-reflect-carnegie-uk-william-perrin?trk=public_profile_article_view>.

[51] William Perrin, *Government online harms*.

[52] William Perrin, *Government online harms*.

[53] William Perrin, *Government online harms*.

Applying legislation in the form of systematic duty of care to social media platforms would provide a framework that would result in consequences if not adhered to.[54] Such a methodology would obligate social media platforms to 'review user content or exercise more control over it.'[55]

The criticism for the UK's introduction of a statutory duty of care is that when enacted, the law sets such a high threshold to satisfy provisions, that it would make it almost difficult to prosecute.[56] When prescribing content based on universal standards for example, objectionable material, the proposed law does not consider the true potential harm thereby hyper-criminalizing actions in the online environment.[57] The operation of free speech may be seriously impacted by the unclear definition of hate speech, which can cause confusion among the general public, social media platforms, OFCOM, and even prosecutors.[58]

Even though the White Paper on Online Harms emphasizes that as it is a fundamental human right to communicate, there may be a positive obligation to intervene voluntarily, in regard to free speech, to regulate online service providers to ensure that the discriminated are protected.[59] The end goal is to achieve a reduction in online harms.

In summary, the UK's approach to a statutory duty of care is plausible and can provide for an additional boost to self-regulation (social media platforms are currently practicing by way creating a conscious duty to act on illegal content). However, there is still a debatable focal point about the definition of harm; should the interpretation of harm be narrow or wide. On its own, it cannot deal with the vigour and complexities of online hate speech or the social media environment in totality.

[54] Daphne Keller, *Broad Consequences*.

[55] Daphne Keller, *Broad Consequences*.

[56] Coe, *Pandora's Box.*

[57] Coe, *Pandora's Box.*

[58] Coe, *Pandora's Box.*

[59] Tambini, *Differentiated Duty of Care,* pp. 28-40.

A STATUTORY DUTY OF CARE ONTO SOCIAL MEDIA PLATFORMS - AUSTRALIA

Since 2010, Australian governments have raised concerns about 'potential unsavoury characters to use the internet as a vehicle for distributing pornography and material of a violent nature to young or otherwise vulnerable individuals'.[60] In March 2010, the Australian Joint Select Committee on Cyber-Safety was established to inquire into 'how young people can be empowered and connect to the Internet, and use new technologies with confidence, knowing that they can use them safely, ethically and with full awareness of risks and benefits.[61]

Responding to the 2019 Christchurch attacks in New Zealand, the Australian government passed new legislation, the Criminal Code Amendment (Sharing of Abhorrent Violent Material Act 2019), that targets ISPs for failure to notify or delete live or streaming violent content.[62] Along with this, the Australian government has implemented reforms regarding child grooming, the regulation of online gambling promotion, the introduction of civil and criminal penalties for the non-consensual sharing of intimate images, and the introduction of the code, aptly named the Australian Code of Practice on Disinformation and Misinformation (the Code), which was developed by an independent body, the Digital Industry Group (DIGI).[63]

In a joint effort to combat online harms, Google, Facebook, Twitter, TikTok and Microsoft signed on a *code* that is governed by Australian legislation.[64]

---

[60] Paula Pyburne, 'Australian Governments and dilemmas in filtering the Internet: juggling freedoms against potential for harm – Parliament of Australia', Parliamentary Library, Parliament of Australia, 8 August 2014. Accessed at: <https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/rp1415/InternetFiltering>.

[61] Pyburne, *Australian Governments and dilemmas in filtering the Internet*.

[62] Parliament of Australia, House of Representatives Select Committee on Social Media and Online Safety, *Report: Social Media and Online Safety,* March 2022. Accessed at: <https://parlinfo.aph.gov.au/parlInfo/download/committees/reportrep/024877/toc_pdf/SocialMediaandOnlineSafety.pdf;fileType%3Dapplication%2Fpdf>.

[63] Select Committee on Social Media and Online Safety, *Social Media and Online Safety*.

[64] Asha Barbaschow, 'Facebook, Google, Microsoft, TikTok, and Twitter adopt Aussie misinformation code', ZDNet Website. Accessed at: < https://www.zdnet.com/article/facebook-google-microsoft-tiktok-and-twitter-adopt-aussie-misinformation-code/>. See also Digital Industry Group, 'DIGI is a nonprofit industry association representing the digital industry in Australia', DIGI Website. Accessed at: <https://digi.org.au/>.

The Code sets out a policy implementation roadmap to regulate the digital environment.[65] The code was developed with principles of protection of freedom of expression where 'the Code gives special attention to international human rights as articulated within the Universal Declaration on Human Rights, including but not limited to freedom of speech'.[66]

Importantly, the Code also offers a definition of *harm* thus providing clarity on what is and what is not harmful to the public.[67] In addition, the Code takes if further by defining and differentiating *misinformation* with *disinformation.*[68] This is beneficial as it provides a clear depiction of the two in the Code. The key difference between misinformation and disinformation is the element of intention. The former being the proliferation of false information regardless of intention to cause harm while the latter, is a deliberate act.

In addition to the Code, Australia's federal parliament enacted the Online Safety Act which came into force on 21st of January 2022, to improve and promote online safety.[69] The Act furnishes existing laws pertaining to online safety making them be more expansive and stronger.[70] Amongst many changes, the Online Safety Act introduces the creation of the role of an eSafety Commissioner to act as a government regulatory agency.[71] This is the first of its kind in the world. The removal of harmful and illegal material is determined by the eSafety Commissioner who has the power to disable access.[72] In taking a holistic approach, the new Act will make it mandatory for online

[65] Commonwealth of Australia, 'Government Response and Implementation Roadmap for the Digital Platforms Inquiry', 12 December 2019. Accessed at: <https://treasury.gov.au/publication/p2019-41708>.

[66] Commonwealth of Australia, *Government Response and Implementation Roadmap*.

[67] Digital Industry Group, 'Australian Code of Practice on Disinformation and Misinformation'. Accessed at: <https://digi.org.au/disinformation-code/>.

[68] Digital Industry Group, *Australian Code,* s3.6.

[69] *Online Safety Act 2021* (Cth) ('Online Safety Act').

[70] eSafety Commissioner, Australian Government, 'Online Safety Act 2021 Fact sheet'. Accessed at: <https://www.esafety.gov.au/sites/default/files/2021-07/Online%20Safety%20Act%20-%20Fact%20sheet.pdf>.

[71] eSafety Commissioner, Australian Government, 'Online Safety Act 2021 takes effect'. Accessed at: <https://www.esafety.gov.au/whats-on/online-safety-act>.

[72] Katharine Gelber, 'A better way to regulate online hate speech: require social media companies to bear a duty of care to users'. *The Conversation,* 14 July 2021. Accessed at: <https://theconversation.com/a-better-way-to-regulate-online-hate-speech-require-social-media-companies-to-bear-a-duty-of-care-to-users-163808>.

platforms to develop new codes. When registered, these codes will in turn make the online industry obligated to act on illegal content.[73]

The Social Media (Anti-Trolling) Bill 2022 was introduced in February to address issues following the High Court decision of *Fairfax Media Publications v Voller*.[74] The Bill essentially 'unmasks' anonymous trolls who post defamatory content on social media. If enacted, this legislation would amongst many other powers, impose liability onto social media platforms by deeming them to be publishers.[75]

However, much of the success of the Australian approach depends upon active compliance by social media publishers. To this end, the Social Media (Anti-Trolling) Bill 2022 may hold important advantages as it aims to eliminate social media networks' ability to assert the innocent distribution defence concerning potentially defamatory content posted by Australian users.[76]

NEW ZEALAND

In New Zealand there are no laws explicitly targeting online hate speech. However, there are a range of other existing laws that have the potential to address some of the harm caused by online hate speech. These include important laws protecting the human rights of New Zealanders, such as the *Bill of Rights Act 1990* (NZ) and *Human Rights Act 1993* (NZ). These laws make it clear that the right to racial equality is protected by law. There are also laws designed to regulate the content of digital communications, such as the *Harmful Digital Communications Act 2015* (NZ) and the *Broadcasting Act 1989* (NZ), both of which aim to put in place standards that reflect

[73] eSafety Commissioner, *Online Safety Act 2021 takes effect*.

[74] *Media Publications v Voller* [2021] HCA 27. See also Parliament of Australia, House of Representatives, Explanatory Memorandum, Social Media (Anti-Trolling) Bill 2022 (Cth). Accessed at: <https://parlinfo.aph.gov.au/parlInfo/download/legislation/ems/r6831_ems_d8a044e1-2ac3-4f15-b90a-7cf5d57b4b2e/upload_pdf/JC004985.pdf;fileType=application%2Fpdf>.

[75] Explanatory Memorandum, Social Media (Anti-Trolling) Bill 2022 (Cth). It should be noted that this Bill lapsed with the proroguing of the Australian Parliament in April 2022.

[76] Business Standard, 'Australia's social media anti-trolling bill raises alarm for tech giants', *Business Standard,* 7 March 2022. Accessed at: <https://www.business-standard.com/article/international/australia-s-social-media-anti-trolling-bill-raises-alarm-for-tech-giants-122030700244_1.html>.

community expectations; however in both cases, the enforcement of these laws has proven insufficient to give rise to effective protection against online hate speech.

As was highlighted in the Christchurch Call, at the Global Internet Forum to Counter Terrorism in Paris, France, that there is a need to take further steps to avoid online harm which include imposing a duty of care onto social media platforms.[77] According to the Helen Clarke Foundation, social media businesses need to invest in and take reasonable steps to prevent harm. This should include strengthening technology-based responses to online hate speech and/or changing their terms of service. The Foundation also recommends the establishment of a regulatory agency to oversee and monitor these social media businesses and impose penalties if they do not take positive action on harm prevention.[78] This regulatory agency should be independent to ensure that compliance of the duty of care are fulfilled by social media platforms.[79] In addition, the Foundation recommends that a suit of powers be bestowed onto this independent regulator for breach of such a duty, including the imposition of substantial fines and personal liability on individual members of senior management.[80]

Royal Commission *Inquiry into The Terrorist Attack on the Muslim Community in Christchurch* established a total of 48 recommendations and on the matter of online hate speech, the recommendations set out improvements to the current legislation.[81] There are existing criminal sanctions for incitement of disharmony on racial grounds. Still, there are no similar protections for hate speech arising from different opinions with regard to religious belief, disability, sexual orientation, or gender identity.[82] The Royal Commission proposes the inclusion of religion, gender, sexuality and disability in

[77] Claire Mason and Kathy Errington, 'Anti-social media: reducing the spread of harm content on social media networks', *Helen Clark Foundation,* 14 May 2019. Accessed at: <https://helenclark.foundation/publications-and-media/anti-social-media/>.

[78] Masson and Errington, *Anti-social media*.

[79] Masson and Errington, *Anti-social media*.

[80] Masson and Errington, *Anti-social media*.

[81] Royal Commission, *Royal Commission of Inquiry into The Terrorist Attack on Christchurch Mosques on 15 March 2019,* New Zealand, 8 March 2022.

[82] David Seymour and Andrew Little, 'Freedom of speech: Do we need to update our Human Rights Act?', *Stuff New Zealand,* 28 June 2019. Accessed at: <https://www.stuff.co.nz/national/politics/opinion/113785976/freedom-of-speech-do-we-need-to-update-our-human-rights-act>.

the protected characteristics, therefore, providing broader protection against wider discriminated groups; adding that the Human Rights Act 1993 should express that 'trans, gender diverse, and intersex people are protected from discrimination.'[83] The hope is to bring about change and reform to the existing framework to include specific groups of people into 'protected categories' in the Act. The Royal Commission of Inquiry further proposes improvements to legislations including the *Human Rights Act 1993* and the *Crimes Act 1961*, to name a few; making these laws fit for purpose by recommending amending legislation to create hate-motivated offences.[84] The recommendations have not been executed by the government; however, it is on its manifesto to ensure that hate speech laws are extended to include more vulnerable groups.[85]

Netsafe, a non-profit organisation that collaborates with the New Zealand government on online safety issues such as education and research, works closely with the Ministry of Justice to provide the public and organisations with information on online safety guidelines and strategies. In addition, Netsafe provides the public with a reporting infrastructure on issues relating to fraud, privacy breaches, online trading complaints, online harassment or bullying and abuse.[86] At present, Netsafe has been developing a voluntary industry code, *Aotearoa New Zealand Code of Practice for Online Safety and Harms*.[87] This Code will establish a self-regulatory framework for the digital industry. Its development is based on a code of practices from other jurisdictions such as the European Union, the United Kingdom and Australia.

[83] Ministry of Justice, New Zealand, 'Proposal against incitement of hatred and discrimination'. Accessed at: <https://www.justice.govt.nz/assets/Documents/Publications/Incitement-Discussion-Document.pdf>.

[84] R*oyal Commission of Inquiry Into The Terrorist Attack On Christchurch Mosques*, Part 10- Recommendations.

[85] Labour 2020 'Our Manifesto To Keep New Zealand Moving'. Accessed at: <https://www.labour.org.nz/policy>.

[86] Netsafe Report, 'Netsafe – Providing free online safety advice in New Zealand'. Accessed at: <https://www.netsafe.org.nz/reportanincident/>

[87] Netsafe 'Aotearoa New Zealand Code of Practice for Online Safety and Harms draft - Netsafe – Providing free online safety advice in New Zealand'. Accessed at: <https://www.netsafe.org.nz/aotearoa-new-zealand-code-of-practice-for-online-safety-and-harms-draft/>.

CONCLUSION

The above comparative analysis suggests that if lawmakers and the broader community are serious about addressing the harm caused by online hate speech, it is critical that we design legislative responses with care.

Statutory models that draw upon common law duties of care owed between manufactures and consumers can be instructive, particularly when used in conjunction with self-regulatory models.[88] The normative impact of these laws can also be enhanced by explicitly describing the nature of harm that can be caused in an online environment, but only if coupled with specific, enforceable statutory rules or standards that set out 'specific targets or quantifiable objectives, (sic) a broader definition of its values and protected groups of individuals'.[89]

In light of the tragedy of the Christchurch Call, New Zealand needs to develop and introduce a statutory duty of care framework to combat online harm. A holistic approach to tackle online harms should be considered, drawing inspiration from the United Kingdom and Australia. Only by adopting an explicit legislative response to online hate speech can lawmakers feel confident that they are taking protective measures on behalf of consumers of social media products.

.

[88] Netsafe, *Aotearoa Code of Practice*.

[89] Tambini, *Differentiated Duty of Care.* p.33.