Bronwyn Howell and Petrus Potgieter

# Navigating the Boundaries of Digital Platform Content Regulation in New Zealand

## Abstract

This article examines the complexities of implementing online content regulation in a small jurisdiction such as New Zealand. Three attempts at hate speech and online content regulation have faltered, in part due to the difficulty of crafting precise legal definitions and different possible conceptions of harm. The 'safer online services and media platforms' policy is the most recent. Given New Zealand's limited market size and the global reach of online platforms, enforcing local content standards is both impractical and potentially ineffective. Most content originates offshore, beyond the scope of domestic legislation, and technological solutions to tailor content to individual user groups are costly and easily circumvented. Existing domestic laws and voluntary industry codes combined with the spillover effects of regulations in larger jurisdictions and international multi-stakeholder efforts likely offer more effective solutions then local legislation. Hence, fostering international cooperation, leveraging global standards and encouraging voluntary compliance should be encouraged.

**Keywords** New Zealand, digital platform content regulation, censorship, free speech, Christchurch Call

New Zealand has a long legacy as a world-leading innovator in telecommunications and internet technology policy, with its 1990s 'light-handed' competition law-based reforms (Howell, 2007) and its government-subsidised nationwide ultrafast broadband network (Heatley and Howell, 2010). More recently, that extended to leadership in internet content moderation policy, via its central role in the Christchurch Call multi-stakeholder network established in 2019 (Wolbers, 2023). Yet so far it has adopted a follower strategy in artificial intelligence (AI) regulation (Ministry of Business, Innovation and Employment, 2024; Collins, 2024) and has had three failed attempts to introduce legislation regulating for hate speech and online platform content moderation, the latest being the coalition government's announcement that it would not be proceeding with legislation introducing a single regulatory agency overseeing both online and physical/broadcast media content (Van Velden, 2024).

This article notes that online content regulation is fraught with definitional difficulties and challenges due to differing perceptions of harm from the same content for different potential end consumers and by extension the groups with which they

**Bronwyn Howell** is a senior lecturer in the School of Management at Victoria University of Wellington and a non-resident senior fellow at the American Enterprise Institute. **Petrus Potgieter** is a professor in the Department of Decision Sciences at the University of South Africa.

are associated (for example, New Zealand as a society). This poses specific difficulties for a small state such as New Zealand (population approximately 5.3 million) when endeavouring to create institutions and legislation to manage both the incidence and the distribution of potentially harmful online content, the vast majority of which is created outside its borders by individuals and firms over which it is unable to exercise formal controls. Ultimately, it must rely on the effects of laws and codes of practice developed elsewhere rather than local laws and practice to moderate the vast majority of online platform content.

To support this assertion, we examine the difficulties faced in New Zealand when endeavouring to define acceptable and unacceptable content legislatively. We reference briefly the difficulties of defining 'hate speech' and 'harm', in legislation proposed in response to the recommendations of the inquiry into the March 2019 Christchurch terrorist attack. The substantive example then examined is the legislation proposed in 2023 to create safer online services and media platforms in New Zealand. Neither of these initiatives has proceeded. We argue that even if passed, the latter proposals would have been largely impotent in influencing online safety in New Zealand. We conclude by suggesting that international multi-stakeholder initiatives such as the Christchurch Call may provide some controls, as do the spillover effects of laws in other, larger jurisdictions to the extent that they can govern the activities of international firms with an online presence in New Zealand. New Zealand initiatives should focus on ensuring that locally produced content meets existing local laws and standards.

### Platform content regulation in New Zealand

The Royal Commission of Inquiry into the Terrorist Attack on Christchurch Mosques on 15 March 2019 reported to the governor-general on 26 November 2020, making specific recommendations regarding hate speech and hate crime-related legislation (Royal Commission, 2020). Two attempts to effect this, in 2021 (Ministry of Justice, 2021) and 2022 (New Zealand Parliament, 2022), failed to proceed, in large part due to difficulties defining exactly what constituted 'hate speech' and the definition of harm. As the interpretation of an utterance and the extent of harm caused are critically dependent on the identity of the recipient(s) (or target) (Small, 2022; Strossen, 2018; Mchangama, 2022), defining hate speech and harm satisfactorily for legislation has proved both difficult and controversial. A third attempt begun in 2023 (Department of Internal Affairs, 2023a) was also abandoned following a change in government at the 2023 general election, with the incoming minister of justice instructing the Law Commission[1] to cease work on hate speech reforms in March 2024 (Manch, 2024) and the minister of internal affairs announcing in May 2024 that the legislation implementing a single platform for digital and physical content moderation would not be progressed (Van Velden, 2024).

Existing provisions were deemed insufficient and inconsistent across platforms, and despite already having legislation to deal with most situations, it was argued, behaviour that is illegal is sometimes tolerated online.

Regulation of online content intertwines with AI regulation because AI tools sit at the core of social media platforms and other applications. AI tools are used to both filter the content uploaded by end users to be viewed by others (Gorwa, Binns and Katzenbach, 2020), and select and order the content distributed to and seen by the end users. From the platform/application perspective, filtering is necessary to ensure that the content displayed does not breach relevant laws in the jurisdictions in which it is viewed, as well as to support the user experience. Platform/application providers may also apply additional screening to filter out content that fails to meet any additional standards they require. For example, a platform operator may add additional conditions, even if they are not legally required, in order to differentiate itself from its rivals or cater to the perceived preferences of its target user audience. Such additional filters have been applied to Google's Gemini text generator to differentiate it from rival OpenAI's ChatGPT (Milmo and Hearn, 2024), for example.

In June 2023, the New Zealand Department of Internal Affairs released proposals for 'safer online services and media platforms' (Department of Internal Affairs, 2023a). The department is responsible for managing the government's content policy and managing relationships with the Classification Office, the independent Crown entity responsible for classifying content available to New Zealanders. The chief censor manages the Classifications Office and exercises its statutory powers. The motivation for the proposals presumed that 'New Zealanders are being exposed to harmful content and its wider impacts more than ever before' (ibid., p.3). Without change, 'New Zealand is at risk of falling behind the protections that other like-minded nations provide' (Department of Internal Affairs, 2023b, p.4). Existing provisions were deemed insufficient and inconsistent across platforms, and despite already having legislation to deal with most situations, it was argued, behaviour that is illegal is sometimes tolerated online. This conclusion was reached despite the apparently very successful blocking of access to the 2019 mosque shooting videos on New Zealand servers by local internet service providers within minutes of the event, and the chief censor making possession of the shooter's manifesto in New Zealand illegal by the next day (Howell, 2019). However, removal of the video and manifesto from overseas servers could not be achieved because New Zealand laws cannot be enforced outside New Zealand's

sovereign boundaries. Similar limitations have been experienced in other jurisdictions, including Australia (Howell, 2024a).

Defining what is harmful is highly content- and consumer-specific. Content that could cause harm or offence to some individuals or groups may not cause harm or offence to others (Strossen, 2018; Mchangama, 2022). Thus, the definition of what constitutes 'safety' comes from consumers' perceptions of a specific piece of content, rather than those of its creator or publisher/host. As platforms cater to many different end consumers, all with different perceptions of what is 'safe' or 'unsafe' content, the decision about whether or not it is acceptable to host it is not simple. Choosing not to host a piece of content because it is not 'safe' for one subset of consumers denies access to it by another subset of consumers for whom it does not constitute a safety risk. Indeed, refraining from hosting it potentially infringes upon the creator's freedom to express their views, the platform host's rights to publish the creator's content and potentially earn revenue from doing so, and the ability of the unharmed consumers to enjoy/benefit from it. And as much as 'society' as a subset of consumers may avoid some collective harms from some potentially harmful content being removed, it is equally potentially harmed by the loss of benefits to those members of society who would not have been harmed, and indeed could have prospered if the content was available.

The difficulty is illustrated by various digital marketing cases brought to the Advertising Standards Authority, the industry self-regulatory body overseeing advertising content in New Zealand (Advertising Standards Authority, n.d.).[2] The vast majority relate to offence taken by specific consumers motivated to lodge a complaint about content they deemed offensive (e.g., transphobic, misogynistic, in bad taste, likely to encourage young children to eat sugary foods), but with which many others would have no problem. Most complaints either were found to have no grounds to proceed or were not upheld. This reflects the very real challenge faced by platforms: the issue to be addressed is frequently not the content per se, but the identity of the user(s) to whom it is displayed. Placing legal responsibility for online safety onto content producers and hosting platforms alone diminishes user actions (and those of their guardians) in monitoring and regulating access to online content and reducing harm (Howell, 2024b). This raises the question of whether the motivation to intervene is primarily to address harm directly linked to specific online content because existing laws and processes have demonstrably failed to ensure reasonable outcomes, or rather to address political concerns – notably, voter perceptions of the government's performance in response to the royal commission on the mosque attacks and pressure from lobby groups. While this question is of interest, we focus here specifically on the content and enforceability of the 2023 safer online services and media platforms proposal.

> … New Zealand's current legal system has powers to deal with 'most awful and illegal content like child sexual exploitation and promotion of terrorism, regardless of whether it is delivered online or through traditional forms of media such as printed publications'

### Safer online services and media platforms

The proposal focused on consumer protection from harmful content in all media types, ranging from 'books, films and radio to social media and blogs and everything in between'. The proposed solution 'introduces more robust consumer protection measures that protects [sic] New Zealanders while maintaining the existing freedoms we enjoy', and came in response to 'widespread concerns about the harm some content is causing children and young people' (Department of Internal Affairs, 2023a, p.3), in addition to concerns about hate speech arising from the 2019 mosque shootings. For example, 83% of respondents to a Classification Office survey expressed concerns about harmful or inappropriate content on social media, video sharing sites and other websites (Classification Office, 2022).

The proposers found that, in principle, New Zealand's current legal system has powers to deal with 'most awful and illegal content like child sexual exploitation and promotion of terrorism, regardless of whether it is delivered online or through traditional forms of media such as printed publications' (Department of Internal Affairs, 2023a, p.4). Hence, no changes to definitions of what is considered legal or illegal in New Zealand were proposed. However, as most of the legislation is over 30 years old, it was not deemed well suited to addressing the specific challenges of an online environment. Existing measures are predominantly reactive, allowing redress only after alleged breaches have occurred and been reported to the relevant authority. Furthermore, they address specific pieces of content only, rather than general categories. Moreover, those responsible for enforcing compliance believed they lacked the tools and powers to oversee online platform activity in respect of online content. A particular concern was that monitoring and oversight is spread among ten separate agencies: government agencies the Department of Internal Affairs, the Classification Office, the New Zealand Police and the New Zealand Customs Service; statutory bodies the Broadcasting Standards Authority and the Film and Video Labelling Body; industry organisations the Advertising Standards Authority, New Zealand Media Council and NZTech; and civil society entities such as Netsafe (see Appendix). This was thought to be leading to confusion among consumers about where to seek information or lodge complaints.

The proposed solution was to create a new independent regulatory (presumably statutory) body, separate from the government, responsible for overseeing all online and other media platform content, with similar online safety obligations to those of Ofcom in the UK.[3] The regulator would 'work with platforms to create a safer environment' and require 'larger or high-risk platforms to comply with codes of practice' developed by industry groups with input from and approval by the regulator (Department of Internal Affairs 2023a, p.6). Platforms will be required to have transparent operating practices in place to meet code requirements. Codes of this form already exist for broadcasters and other traditional media, currently overseen by bodies such as the Broadcasting Standards Authority, the New Zealand Media Council and the Advertising Standards Authority. However, these would have had to be renegotiated, alongside the negotiation of new codes for social media and other internet platforms, to ensure that consistent standards apply across all media.

The regulator would have had powers to 'check information from platforms to make sure they follow the codes and could issue penalties for serious failures of compliance' and 'powers to require illegal material to be removed quickly from public availability in New Zealand' (ibid., p.7). The regulator would thus enforce existing censorship legislation, in addition to exercising new powers 'to deal with material that is illegal for other reasons, such as harassment or threats to kill' (ibid.). To do this, the regulator would require the regulated platforms to submit transparency reports, review complaints, share information with domestic and international agencies, submit regulated platforms to periodic and ad-hoc audits, and request relevant information from the regulated platforms as necessary. Enforcement powers would include directions to take remedial action to address identified gaps or deficiencies, formal warnings, civil penalties for non-compliance, and directions to remove non-compliant material in a stipulated time frame.

Feedback was sought from the public on what other kinds of illegal material the regulator should have powers to deal with. Submission summaries were released in May 2024, along with the minister's announcement that the legislation would not proceed.

To the extent that any New Zealand-specific content warrants special treatment, attending to this via specific legislation empowering the relevant bodies already in place rather than relying upon new online codes would appear to be a more robust resolution.

### Discussion

While there may be some merit in coordinating the many bodies with a stake in content oversight and consistently reporting their activities,[4] it begs the question of why additional powers of censorship are warranted, or can even be effectively enforced, given that most of the content viewed in New Zealand is created offshore by companies outside the jurisdiction of the New Zealand legal system.

First, to the extent that existing codes already address current legal requirements, albeit customised for specific circumstances, no further additional codes appear to be necessary. There are many examples of adherence to codes created and overseen by independent entities which have the effect of providing information to consumers with higher standards than legally required, in both the New Zealand and international contexts. For example, Meta finances a quasi-judicial oversight board with an independent membership appointed from public nominations, via a system managed by US law firm Baler McKenzie, to oversee content on Facebook and Instagram to ensure that these platforms promote free expression by making principled decisions on content hosting according to their transparent policies (Wong and Floridi, 2023). The board's independence and wide scope is asserted because Meta does not control the appointments process; the current board has membership from across the globe, albeit with a US bias (six of 18 members) (Meta Oversight Board, 2024). Likewise, civil society organisation Netsafe has developed a voluntary code for platforms operating in New Zealand, with input from Meta, Google, TikTok, Twitch and Twitter (now X) (Netsafe, n.d.). Similar initiatives have taken place internationally.

As New Zealand is so small relative to the international content providers, with only 0.05% of global population, relying on existing measures rather than creating yet another set of codes and adding additional layers of bureaucracy appears prudent. To the extent that any New Zealand-specific content warrants special treatment, attending to this via specific legislation empowering the relevant bodies already in place rather than relying upon new online codes would appear to be a more robust resolution. However, Ververis et al. (2024) have demonstrated that even website blocklists (or instructions) issued by authorities at national level in the EU, for example, are imperfectly put into practice and that there is a lack of transparency. So the simple act of legislating does not guarantee a more effective outcome.

Second, while New Zealand exists as a sovereign state capable of making its own laws, the entities engaged in creating and hosting content viewed in New Zealand are increasingly neither physically nor legally present in the country.[5] The extent to which they understand or can be held to account to a New Zealand code (or even current New Zealand law) is debatable. This is illustrated by a recent example involving the legality of online advertisements for a foreign betting agency involving a New Zealand sportsman

(Howell, 2023). Whereas the Department of Internal Affairs, having direct regulatory oversight of gambling and responsibility for relationships with the Advertising Standards Authority, asserted that the advertisements were not in breach of New Zealand law, Facebook erroneously took them down on the mistaken understanding that the New Zealand legal provisions were the same as those in Australia. Similarly, the Australian eSafety commissioner was found to be legally unable to force X to take down content internationally deemed unacceptable in Australia.[6] Whether other platforms' (Meta, TikTok, Google) actions in taking down the content internationally was due to the eSafety commissioner's order alone or because of obligations as signatories to voluntary international codes such as the Christchurch Call is arguable (Howell, 2024a).

The New Zealand market is such a small share of the global custom for many platforms that they have few incentives to understand (or may place a low priority on) the nuances of a New Zealand-specific code or law. For the most part, compliance with content codes relies upon AI and not human decision making. If New Zealand code requirements are sufficiently different from those applying in other jurisdictions, then large international platforms are likely to find it extremely difficult to manage the 'safety' of New Zealand consumers by focusing only on the nature of the content alone. The identity, location and preferences/restrictions of the content viewer will need to be known before an effective decision can be made about whether it is 'safe' to let that individual view specific content. This does not mean that there is not a role to play in New Zealand in ensuring that locally produced content obeys local laws. Rather, it means that the focus of compliance and harm avoidance lies in the relationship between the government and content creators that can be monitored and enforced, rather than with international online media platforms, where enforceability is extremely limited.

Yet a third issue arises due to technical considerations. Providers can moderate content based on a user's IP address and browser information (e.g., 'cookies'). The first is used to prevent content distribution in specific geographic locations, the latter to distinguish between different individuals using the same address. However, these can be masked by users to get access to content that otherwise may have been withheld (e.g., via VPNs, anonymising browsers and new accounts). A third party with access to a data stream (e.g., a broadband provider or a government entity intercepting traffic at a key point) could block all content from a specific server, but not individual items communicated using Hypertext Transfer Protocol Secure (HTTPS) (very commonly used), due to encryption. Hence, direct onshore reviewing and blocking of selected traffic generated offshore is not an option open to a country like New Zealand, where residents expect and need access to the usual panoply of online services available elsewhere (Howell and Potgieter, 2023; Howell, 2024a). Interception of this kind is possible in China because the country has developed its own onshore services that serve as partial substitutes (Wu et al., 2023).

As the prime minister's chief science advisor has observed: 'As a relatively small economy, NZ doesn't have the market power to incentivise suppliers to comply with overly onerous regulation that is not in place elsewhere'

In sum, therefore, there is negligible evidence that existing provisions, including existing laws and all statutory, industry-led and multi-stakeholder codes, are ineffective. Confusion about where to file complaints could be satisfactorily addressed by publicity and education about the existing arrangements, and even a single portal linking a consumer directly to the relevant separate enforcement entities. Even if new laws were passed, the ability to enforce them effectively is limited, so it appears little will be lost if New Zealand adopts a follower strategy for internet content moderation in addition to AI regulation. Indeed, if international endeavours such as civil society initiatives, or 'Brussels effect'[7] regulations in larger jurisdictions where platform companies operate, address safety concerns satisfactorily for the vast majority of the content New Zealanders view, or the AI applications they use, then no extra regulation may be necessary at all. Existing codes may continue to prove sufficient.

## Conclusion

When New Zealand took an innovative approach to telecommunications policy in the 1980s, the subjects of the ensuing regulation were firms with a defined legal existence and presence within New Zealand, creating and providing services within New Zealand and almost exclusively to New Zealanders. Governing their activities under New Zealand law was straightforward, even though many of the firms entering the local market were foreign (e.g., Bell Atlantic, BellSouth, Vodafone, Econet). When it comes to the regulation of online content, while the consumers of internet content in New Zealand may have valid concerns about New Zealand-specific policy and legislation, the creators, hosts and promulgators of the vast majority of the applications, their outputs and contents will not be within the scope and jurisdiction of New Zealand laws. As the prime minister's chief science advisor has observed: 'As a relatively small economy, NZ doesn't have the market power to incentivise suppliers to comply with overly onerous regulation that is not in place elsewhere'.

In this context, New Zealand's leadership of the Christchurch Call takes on a new significance (Wolbers, 2023). New Zealand assumed this role not due to legislative experience but because, by dint of fate, it was in the international spotlight at the 'right time'. The country led by advocacy, not by example, when the Call

was founded because the massacre occurred in a country led by a prime minister with a strong international profile and the right connections to bring targeted stakeholders to the table. New Zealand stands to benefit from the Call initiatives, at least in relation to terrorist content, regardless of whether or not they take form in local laws, so long as the self-regulatory processes of the Call are effective. This is one reason why the New Zealand government underwrote the Call's administrative functions until the end of June 2024.[8]

Small countries such as New Zealand must rely in large part on international initiatives if they are to influence the behaviour of international media platform firms. But as Wolbers (2024) indicates, these initiatives must be self-sustaining and independent from too much government as well as industry influence if the codes of practice are to be effective. This is a hard task to achieve. If the Christchurch Call and other civil society initiatives targeting other forms of objectionable online content cannot succeed, then New Zealand and other small countries must continue to rely only on spillover effects from laws passed in other countries influencing the behaviour of the international platforms making their content available to New Zealand citizens as members of a global audience.

1 The Law Commission is an independent authority, like the privacy commissioner and the Human Rights Commission, established under the Law Commission Act 1985 and accountable to Parliament to review laws and make recommendations to government on how they can be improved.

2 Unlike the similarly-named UK Advertising Standards Authority, the New Zealand entity is not a statutory body.

3 Established in 2003, by the Office of Communications Act 2002, Ofcom has additional responsibilities in regulating telecommunications, broadcasting and postal services not envisaged in the New Zealand proposal.

4 There is no apparent efficiency benefit from the process, as it is likely that all of the existing codes would have remained in some form, overseen by the respective industry experts, in addition to the overlay of regulatory obligations for the firms and the new regulator. If what was sought was ease of access for consumers to the existing processes (which it has not been established were not working), then arguably this could have been achieved with simply a single portal linking to each of the specific complaints processes. If what was intended was a single process for all media types, then inevitably there would have been compromises in effectiveness, as it is far from clear that one code with a generic set of criteria would suffice given the various different objectives for different forms of content (e.g., advertising content is not assessed against the same criteria as broadcast entertainment).

5 While Netsafe stakeholders Meta, Google, Twitch and X have a legal presence in New Zealand, this is principally to manage their advertising and other commercial interests; content moderation activity occurs outside New Zealand.

6 *eSafety Commissioner v X Corp* [2024] FCA 499 (https://www.judgments.fedcourt.gov.au/judgments/Judgments/fca/single/2024/2024fca0499). An interim injunction was granted while the judge considered the case, meaning the content was not available anywhere in the world for some days. However, the judge eventually found in favor of X and the content was made available outside Australia.

7 The Brussels effect refers to the ability for EU laws to apply to firms and markets outside the EU, by a combination of legislative specifications capturing any entity outside the EU trading with an EU resident, and advocacy for early-mover EU laws to become a standard for regulation across the globe (Bradford, 2012).

8 Funding was withdrawn due to a combination of financial pressures and a change of government leading to different political priorities.

## References


Advertising Standards Authority (n.d.) 'Search/browse decisions, https://www.asa.co.nz/decisions/search-browse-decisions/

Bradford, A. (2012) 'The Brussels effect', *Northwestern University Law Review*, 107 (1), https://scholarlycommons.law.northwestern.edu/nulr/vol107/iss1/1/

Classification Office (2022) *What We're Watching: New Zealanders' views about what we see on screen and online*, https://www.classificationoffice.govt.nz/media/documents/What_Were_Watching_Accessible.pdf

Collins, J. (2024) 'Approach to work on artificial intelligence', Cabinet paper, 26 June, https://www.mbie.govt.nz/dmsdocument/28913-approach-to-work-on-artificial-intelligence-proactiverelease-pdf

Department of Internal Affairs (2023a) 'Safer online services and media platforms', discussion document, https://www.dia.govt.nz/diawebsite.nsf/Files/online-content-regulation/$file/Safer-Online-Services-and-Media-Platforms-Discussion-Document-June-2023.pdf

Department of Internal Affairs (2023b) 'Safer online services and media platforms: executive summary', https://www.dia.govt.nz/diawebsite.nsf/Files/online-content-regulation/$file/Safer-Online-Services-and-Media-Platforms-Executive-Summary-June-2023.pdf

Gorwa, R., R. Binns and C. Katzenbach (2020) 'Algorithmic content moderation: technical and political challenges in the automation of platform governance', *Big Data and Society*, 7 (1), 2053951719897945, https://doi.org/10.1177/2053951719897945

Heatley, D. and B. Howell (2010) 'Structural separation and prospects for welfare-enhancing price discrimination in a new "natural monopoly" network: comparing fibre broadband proposals in Australia and New Zealand', Wellington: New Zealand Institute for the Study of Competition and Regulation, https://ir.wgtn.ac.nz/items/aa067739-ae99-41dd-aa4b-4abcd61dfdcb

Howell, B. (2007) 'A pendulous progression: New Zealand's telecommunications regulation 1978–2007', New Zealand Institute for the Study of Competition and Regulation, https://researcharchive.vuw.ac.nz/xmlui/bitstream/handle/10063/3974/Pendulous_Progress_v_4_12_Nov.pdf?sequence=1

Howell, B. (2019) 'Gunning for Facebook and free speech', *AEIdeas*, 26 March, https://www.aei.org/technology-and-innovation/gunning-for-facebook-and-free-speech

Howell, B. (2023) 'The tyranny of the comments section: unforced errors in online content moderation down under', *AEIdeas*, 19 April, https://www.aei.org/technology-and-innovation/the-tyranny-of-the-comments-section-unforced-errors-in-online-content-moderation-down-under/

Howell, B. (2024a) 'Australia rules social media, O.K?', 20 May, *AEIdeas*, https://www.aei.org/technology-and-innovation/australia-rules-social-media-ok/

Howell, B. (2024b) 'Who is ultimately responsible for children's internet safety?', *AEIdeas*, 2 February, https://www.aei.org/technology-and-innovation/who-is-ultimate-responsible-for-childrens-Internet-safety/

Howell, B. and P. Potgieter (2023) 'Shooting the messenger: an evaluation of New Zealand's safer online services and media platforms proposal', paper presented at the International Telecommunications Society 16th Asia-Pacific regional conference, 27 November

Manch, T. (2024) 'Hate speech work cancelled, but hate crime work continues', *Post*, 19 March, https://www.thepost.co.nz/politics/350217913/hate-speech-work-cancelled-hate-crime-work-continues

Mchangama, J. (2022) *Free Speech: a history from Socrates to social media*, Basic Books

Meta Oversight Board (2024) 'Get to know our board', https://www.oversightboard.com/meet-the-board/

Milmo, D. and A. Hern (2024) '"We definitely messed up": why did Google AI tool make offensive historical images?', *Guardian*, 8 March, https://www.theguardian.com/technology/2024/mar/08/we-

definitely-messed-up-why-did-google-ai-tool-make-offensive-historical-images

Ministry of Business, Innovation and Employment (2024) 'Digital policy', https://www.mbie.govt.nz/business-and-employment/economic-growth/digital-policy

Ministry of Justice (2021) 'Proposals against incitement of hatred and discrimination', discussion document, https://www.justice.govt.nz/assets/Documents/Publications/Incitement-Discussion-Document.pdf

Netsafe (n.d.) 'Our story', https://netsafe.org.nz/aboutnetsafe/our-story/

New Zealand Parliament (2022) 'Human Rights (Incitement on Ground of Religious Belief) Amendment Bill', https://www.parliament.nz/en/pb/sc/make-a-submission/document/53SCJU_SCF_BILL_130197/human-rights-incitement-on-ground-of-religious-belief

Office of the Prime Minister's Chief Science Advisor (2023) 'Regulating risks associated with AI', https://bpb-ap-se2.wpmucdn.com/blogs.auckland.ac.nz/dist/f/688/files/2023/07/Regulatory-landscape.pdf

Privacy Commissioner (2024) 'New Zealanders more concerned than excited by AI', 13 March, https://www.privacy.org.nz/publications/statements-media-releases/new-zealanders-more-concerned-than-excited-by-ai/

Royal Commission of Inquiry into the Terrorist Attack on Christchurch Mosques on 15 March 2019 (2020) *Hate Speech and Hate Crime Related Legislation*, https://christchurchattack.royalcommission.nz/assets/Publications/Hate-speech-and-hate-crime-related-legsilation-Companion-legal-paper.pdf

Small, Z. (2022) 'How confusion over hate speech law prompted rethink: Prime Minister Jacinda Ardern vows to "get the change right"', *Newshub*, 22 March, https://www.newshub.co.nz/home/politics/2022/03/how-confusion-over-hate-speech-law-prompted-rethink-prime-minister-jacinda-ardern-vows-to-get-the-change-right.html

Strossen, N. (2018) *Hate: why we should resist it with free speech, not censorship*, Oxford University Press

Van Velden, B. (2024) 'Brooke van Velden MP's post', Facebook, 14 May, https://www.facebook.com/100063738000728/posts/947408990727026/?mibextid=QwDbR1andrdid=kU1VjQy26uiGTzLR

Ververis, V., L. Lasota, T. Ermakova and B. Fabian (2024) 'Website blocking in the European Union: network interference from the perspective of Open Internet', *Policy and Internet*, 16 (1), pp.121–48

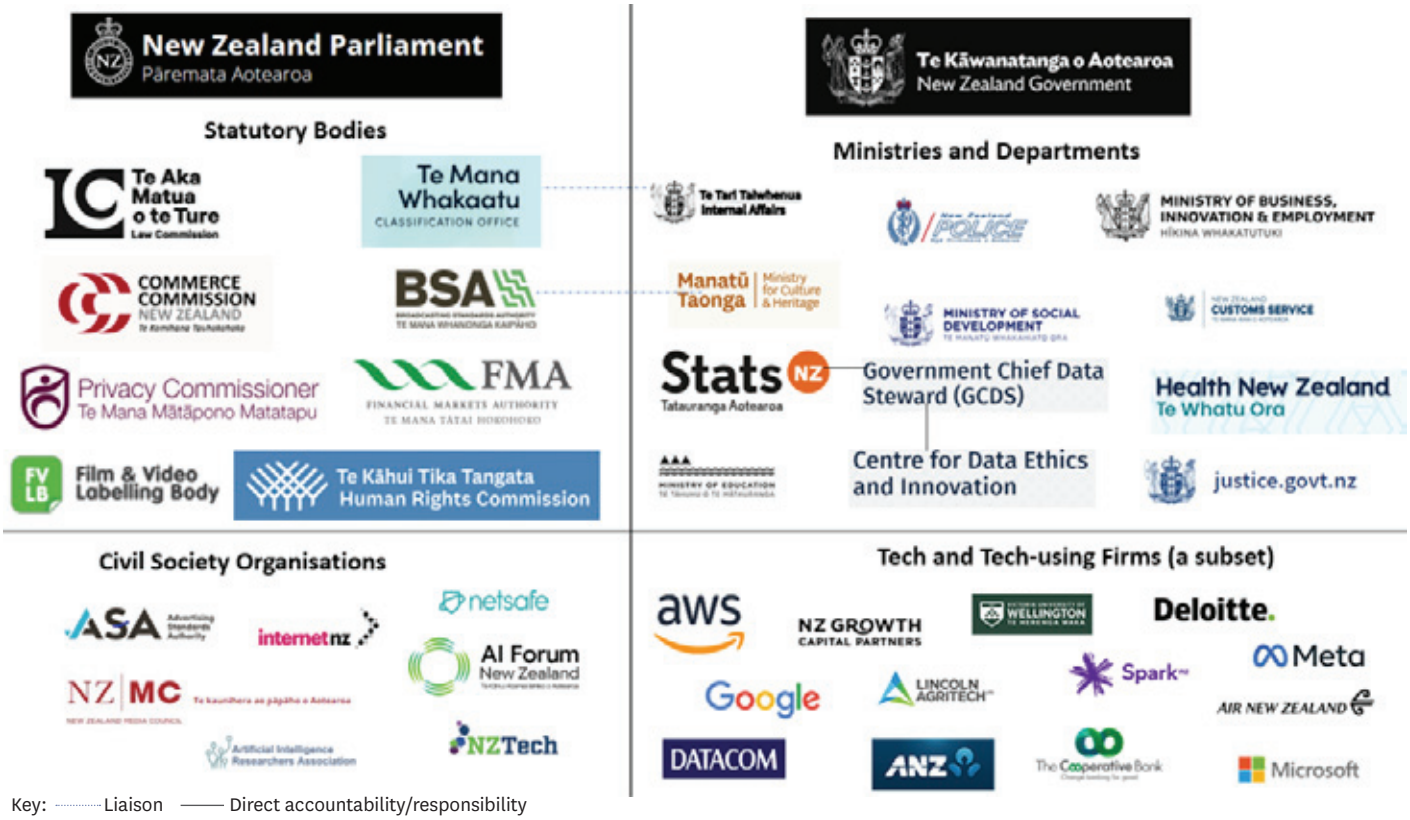
Wolbers, R. (2023) 'The future of the Christchurch Call to Action: how to build multistakeholder initiatives to address content moderation challenges', Ian Axford Fellowship report, https://axfordfellowships.org.nz/wp-content/uploads/2023/08/2023-08-Rachel-Wolbers-Axford-Fellow-The-Future-of-the-Christchurch-Call-to-Action.pdf

Wolbers, R. (2024) 'The Christchurch Call to Action loses its government funding: that's a good thing', *Tech Policy Press*, 3 July, https://www.techpolicy.press/the-christchurch-call-to-action-loses-its-government-funding-thats-a-good-thing/

Wong, D. and L. Floridi (2023) 'Meta's oversight board: a review and critical assessment', *Minds and Machines*, 33, pp.261–84, https://doi.org/10.1007/s11023-022-09613-x

Wu, M., J. Sippe, D. Sivakumar, J. Burg, P. Anderson, X. Wang, K. Bock, A. Houmansadr, D. Levin and E. Wustrow (2023) 'How the great firewall of China detects and blocks fully encrypted traffic', paper presented at the 32nd USENIX Security Symposium, https://www.usenix.org/conference/usenixsecurity23/presentation/wu-mingshi

# Appendix: the New Zealand AI institutional environment




Key: ····· Liaison —— Direct accountability/responsibility