THE Helen Clark FOUNDATION

Anti-social media:
reducing the spread of harmful content on social media networks

CLAIRE MASON
KATHERINE ERRINGTON

CONTENT

Auckland University of Technology is partnering with The Helen Clark Foundation while the Foundation becomes established. This aligns with the University's commitment to research for the public good, and public communication of research. The University supports the Foundation's aim to fund and facilitate research on a range of topics; however its support is not an endorsement of any individual item of work, output or statement made by the Foundation's staff, directors or patron.

ABOUT THE HELEN CLARK FOUNDATION

The Helen Clark Foundation is an independent public policy think tank based in Auckland, at the Auckland University of Technology. It is funded by members and donations. We advocate for ideas and encourage debate, we do not campaign for political parties or candidates. Launched in March 2019, the foundation issues research and discussion papers on a broad range of economic, social and environmental issues.

OUR PHILOSOPHY

New problems confront our society and our environment, both in New Zealand and internationally. Unacceptable levels of inequality persist. Women's interests remain underrepresented. Through new technology we are more connected than ever, yet loneliness is increasing, and civic engagement is declining. Environmental neglect continues despite greater awareness. We aim to address these issues in a manner consistent with the values of former New Zealand Prime Minister Helen Clark, who serves as our patron.

OUR PURPOSE

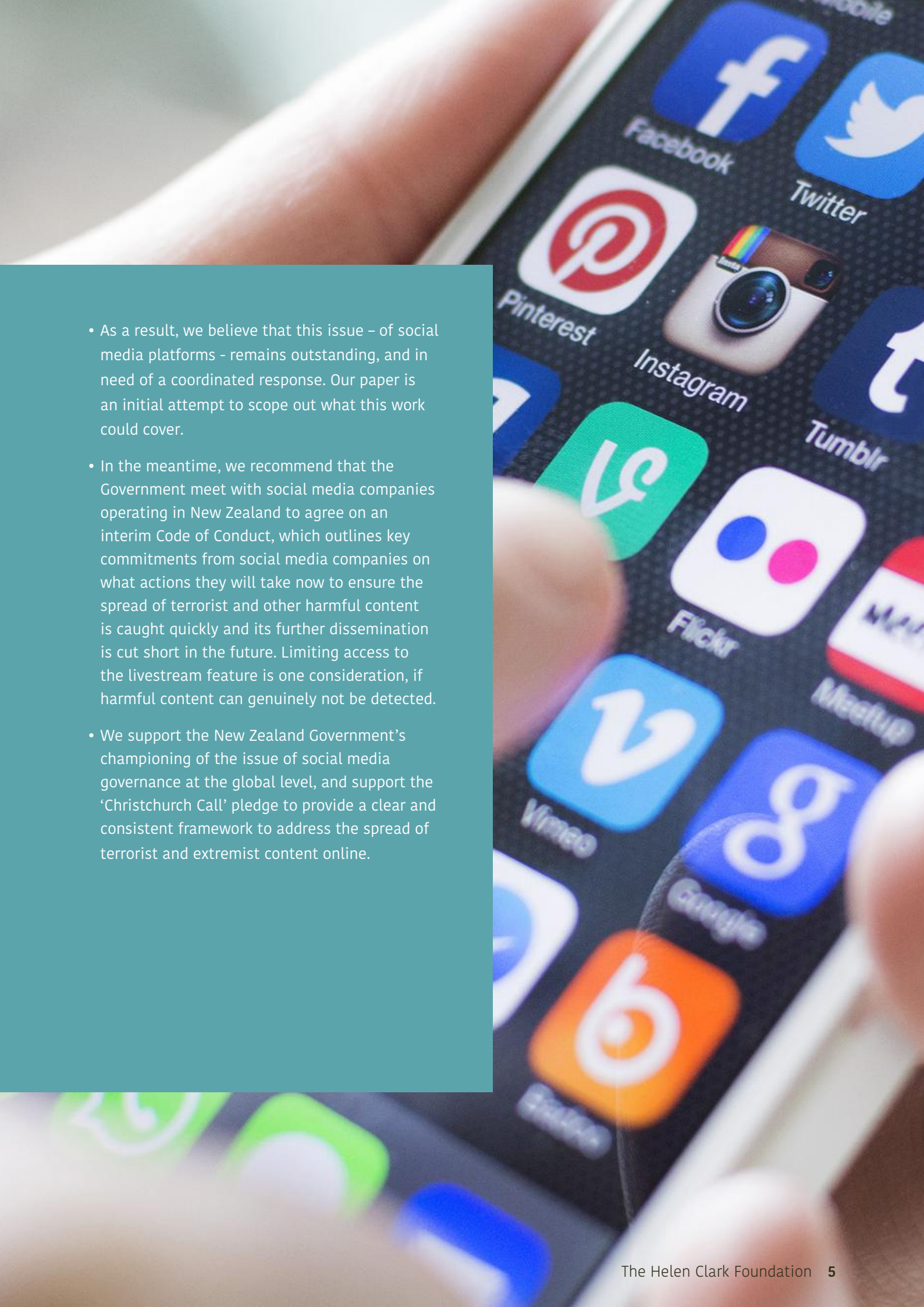
The Foundation publishes research that aims to contribute to a more just, sustainable and peaceful society. Our goal is to gather, interpret and communicate evidence in order to both diagnose the problems we face and propose new solutions to tackle them. We welcome your support, please contact **director@helenclark.foundation** for more information about getting involved.

EXECUTIVE SUMMARY

Anti-social media: reducing the spread of harmful content on social media networks

- In the wake of the March 2019 Christchurch terrorist attack, which was livestreamed in an explicit attempt to foster support for white supremacist beliefs, it is clear that there is a problem with regard to regulating and moderating abhorrent content on social media. Both governments and social media companies could do more.
- Our paper discusses the following issues in relation to what we can do to address this in a New Zealand context; touching on what content contributes to terrorist attacks, the legal status of that content, the moderation or policing of communities that give rise to it, the technical capacities of companies and police to identify and prevent the spread of that content, and where the responsibilities for all of this fall - with government, police, social media companies and individuals.
- We recommend that the New Zealand Law Commission carry out a review of laws governing social media in New Zealand. To date, this issue is being addressed in a piecemeal fashion by an array of government agencies, including the Privacy Commission, the Ministry of Justice, the Department of Internal Affairs, and Netsafe.
- Our initial analysis (which does not claim to be exhaustive) argues that while New Zealand has several laws in place to protect against the online distribution of harmful and objectionable content, there are significant gaps. These relate both to the regulation of social media companies and their legal obligations to reduce harm on their platforms and also the extent to which New Zealand law protects against hate speech based on religious beliefs and hate-motivated crimes.
- The establishment of the Royal Commission into the attack on the Christchurch Mosques on 15 March 2019 (the Royal Commission)[1] will cover the use of social media by the attacker. However the Government has directed the Royal Commission not to inquire into, determine, or report in an interim or final way on issues related to social media platforms, as per the terms of reference.

1 For more information on the Royal Commission, please see https://www.dia.govt.nz/Royal-Commission-of-Inquiry-into-the-Attack-on-Christchurch-Mosques



- As a result, we believe that this issue – of social media platforms - remains outstanding, and in need of a coordinated response. Our paper is an initial attempt to scope out what this work could cover.
- In the meantime, we recommend that the Government meet with social media companies operating in New Zealand to agree on an interim Code of Conduct, which outlines key commitments from social media companies on what actions they will take now to ensure the spread of terrorist and other harmful content is caught quickly and its further dissemination is cut short in the future. Limiting access to the livestream feature is one consideration, if harmful content can genuinely not be detected.
- We support the New Zealand Government's championing of the issue of social media governance at the global level, and support the 'Christchurch Call' pledge to provide a clear and consistent framework to address the spread of terrorist and extremist content online.

RECOMMENDATIONS

- We recommend a legislative response is necessary to address the spread of terrorist and harmful content online. This is because ultimately there is a profit motive for social media companies to spread 'high engagement' content even when it is offensive, and a long standing laissez faire culture inside the companies concerned which is resistant to regulation.

Specifically we recommend the New Zealand Government:

- Direct the New Zealand Law Commission to review regulation of social media. The current legislative landscape is a patchwork of legislation much of which predates social media.
- Establishes an independent regulatory body to oversee social media companies in New Zealand. The New Zealand Media Council and the Broadcasting Standards authority provide a basis for how such an agency could be structured.
- Imposes a statutory duty of care on social media companies. Social media companies would need to invest in and take reasonable measures to prevent harm by, for example, improving their technology-based responses or make changes to their terms of service; otherwise they would face penalties from a regulatory body mandated to oversee and monitor online harms.
- Carefully considers how hate speech and hate crimes are currently protected and prosecuted against under New Zealand law. The current definition which is limited to 'racial disharmony' is too narrow and fails to capture hate speech directed at religious groups, gender and LQBTI+ individuals. Until March 15, the most recent successful prosecution for hate speech was in the 1970s. While the bar *should* be high, this suggests it is too high.
- Meets with social media companies operating in New Zealand to agree on an interim plan of action, similar to the EU's Code of Conduct on Countering Illegal Hate Speech Online or the UK's Digital Charter, which includes commitments from social media companies to adapt their current processes and policies, including their detection and removal procedures, to ensure the spread of terrorist and other harmful content is caught quickly and its further dissemination is cut short.
- Directs New Zealand's intelligence services to develop a high-level strategy outlining their commitments to combatting white supremacist and far right extremism and what steps they will take to prioritise this issue, and make this document public.
- Continues to champion the issue of social media governance at the global level – such as the 'Christchurch Call' Summit in May 2019 - to ensure a multi-jurisdictional approach to addressing the spread of terrorist and harmful content online is prioritised.

INTRODUCTION

On 15 March 2019, a gunman opened fire in two mosques in Christchurch, New Zealand, killing 50 people and injuring 50 more. During his attack at the Al Noor mosque, where the majority of victims were killed, the alleged perpetrator livestreamed his actions directly on Facebook via a helmet-mounted camera.

As later confirmed by Facebook,[3] the livestreaming of the Christchurch terrorist attack did not trigger its current monitoring mechanisms and it was not until a user alerted Facebook to the video – 29 minutes after livestreaming of the attack started and 12 mins after it ended – that it become aware of the issue.[6] By that point approximately 4,000 people had already viewed the video.

It was then widely shared on Facebook, quickly replicated and shared on other platforms, including YouTube and Twitter and appeared on several news media outlets. Within the first 24 hours of the terrorist attacks, Facebook removed more than 1.5 million uploads of the video.

Facebook's delay in disabling the livestream video of the attack on its platform and quickly preventing the further uploading and dissemination of the video has thrown a spotlight on the capacity and willingness of social media platforms to rapidly and effectively respond to terrorist and harmful content online, both in New Zealand and globally.

Many questions are now being asked about why Facebook was so slow to act, what more could social media companies have done and what decisive action needs to be taken to restrict the livestreaming of extremist and violent content in the future and hold social media companies to account.

This paper intends to contribute to the discussion by outlining what actions can be taken in the short term by the New Zealand Government.

Image credit: James Dann.

2 https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/

3 https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/

IT TOOK FACEBOOK 29 MINUTES TO RESPOND TO THE LIVE-STREAMED ATTACK: WHY THE DELAY

Facebook's sluggish reaction in detecting and removing the livestream video of the Christchurch terrorist attack demonstrates the distinct challenges social media companies face in effectively controlling the appearance of terrorist and harmful content on their platforms.

As is standard practice, most social media platforms have monitoring mechanisms in place that utilise both artificial intelligence and human analysis to filter out harmful content online. As conceded by Facebook,[4] and generally acknowledged across the sector,[5] these systems have their limits and need to continually adapt in the face of new threats.

While social media companies have worked hard to develop monitoring systems that pick up other harmful content, such as suicide, child pornography and graphic violence, Facebook asserts thats it is difficult to detect the livestreaming of mass murder. It noted in the aftermath of the Christchurch terrorist attack that artificial intelligence monitoring systems need to be fed large amounts of harmful content so that they can 'learn' how to detect similar content online.

Due to a lack of content comparable to the Christchurch terrorist attack, as well as to the proliferation of visually-similar online video gaming content that confuses artificial intelligence monitoring systems, Facebook says it is challenging to pick up real-life murder online.[6] Once they are aware of harmful content, Facebook and other social media companies do have the technology to extract a 'fingerprint' from it (known as hashing), which enables them to quickly detect and remove harmful content when it appears elsewhere on their platforms.

As we saw in the wake of the Christchurch terrorist attack, harmful content can be easily edited or reformatted to change the hashing fingerprint – for example, the Global Internet Forum to Counter Terrorism reported it found 800 "visually distinct videos" of the attack[7] - making it much harder for social media platforms to pick up and curb the spread of harmful content.

Developments in the business model of social media networks have also seen internet companies continually trying to find new ways to meet their users' need to interact and communicate instantaneously online. Facebook has championed its own live-streaming service for this specific purpose, but considering how susceptible it is to harmful use, the nature of its 'live' function needs to be re-evaluated.

Facebook confirmed it would look at how to strengthen the rules for using Facebook Live, including whether to impose restrictions on who can use the service[8] but has to date refused to make any substantive changes to its livestreaming service.[9]

At the same time, as Facebook has admitted, its internal policies – in particular, its "acceleration process" – also thwarted the rapid detection

4 https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/
5 https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis/
6 https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/
7 https://www.gifct.org/press/industry-cooperation-combat-violent-extremism-all-its-forms/
8 https://www.nzherald.co.nz/business/news/article.cfm?c_id=3&objectid=12217454
9 https://www.newshub.co.nz/home/new-zealand/2019/04/facebook-won-t-put-delays-on-livestreams-mark-zuckerberg.html

and removal of the Christchurch terrorist attack video as it is was only primed to accelerate the assessment of videos containing potential suicide content to human reviewers for immediate action.[10] While unclear at this stage, it is also possible that social media platforms' algorithms inadvertently pushed the video of the Christchurch terrorist attack to like-minded users, enabling it to be more rapidly uploaded and shared.

The scale of online content being generated and the capacity for social media platforms to monitor it effectively also needs to be acknowledged. Facebook, the world's largest social media platform, has more than 2.2 billion active users each month with YouTube following close behind with 1.9 billion. In 2018, more than 300 hours of video were uploaded to YouTube every minute, and 500 million tweets are sent on Twitter every day.[11] In an attempt to better monitor the vast amount of content on their platforms, Facebook and YouTube have invested in the improvement of their automated systems and hired tens of thousands of people around the world to assist with content moderation.

It must be noted, however, that there is no commercial incentive for Facebook, and other social media companies with similar business models, to self-regulate content effectively on their platforms. While it is not in Facebook's business interests to host extreme and harmful content, hosting 'engaging' content is central to its business model as it is the key driver of its advertising revenue. 'Borderline' content, or content which is sensationalist or provocative in nature, creates more user engagement - the more users 'like', share and comment on content, the more information Facebook can collect on users and the more targeted service it can provide to digital advertisers. Digital advertising is big business - Google and Facebook dominate the market with Facebook's net revenue from global digital advertising forecasted to hit over US$67 billion in 2019.[12]

Social media companies – in allowing their users to generate more 'engaging' content in pursuit of commercial ends – have played a key role in facilitating an environment where 'borderline' content progressively crosses the line.

Considering how central 'borderline' content is to Facebook's commercial interests, it has a clear conflict of interest in ensuring the effective regulation of harmful content on its platform. As will be discussed in more detail below, social media companies should no longer be allowed to regulate themselves. Rather a counterweight needs to put in place – such as significant financial penalties or corporate liability – to incentivise social media companies to address harmful content effectively on their platforms despite the potential impact such measures may have on the proliferation of 'borderline' content and the lucrative advertising revenues it brings.

10 https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/

11 https://www.omnicoreagency.com/facebook-statistics/; https://www.omnicoreagency.com/youtube-statistics/; https://www.omnicoreagency.com/twitter-statistics/

12 https://www.emarketer.com/content/global-digital-ad-spending-2019

BEYOND TECHNICAL SOLUTIONS – WHAT MORE CAN BE DONE?

While developing better systems to catch harmful content online is essential, a wider regulatory approach designed to combat the spread of harmful content online comprehensively, and which includes an obligation on social media companies to continually invest in developing more advanced technical solutions, must be prioritised.

As is clear from the Christchurch terrorist attack, while social media has the power to bring about positive changes in society, it can also be used to facilitate significant harm and risk. As has been repeatedly called for since the Christchurch terrorist attack, the time has come for social media companies to be subject to greater regulation and higher standards.

As recently indicated by Facebook's founder and CEO, Mark Zuckerberg, social media companies are likely to welcome a more active role from governments and regulators as the responsibility for monitoring harmful content is too great for social media companies alone.[13] Currently, social media platforms that operate in New Zealand are subject to a patchwork of regulation including at least five agencies – the Ministry of Justice, the Department of Internal Affairs, Netsafe, the Privacy Commission, and the Police.[14] As is standard practice globally, social media companies self-regulate when it comes to moderating content on their platforms through tools such as terms of service and community standards to keep a check on user behaviour and, as described above, automated monitoring systems and human content moderators to catch potentially harmful content and determine whether it should be removed or not.

International approaches to regulating terrorist and harmful content online

With the rise of harmful content online and its potential impact on the public interest, governments in several countries have taken a more interventionist approach by introducing legislation or regulations with strict deadlines and high penalties for non-compliance in order to oblige social media platforms to rapidly remove harmful content.

For example, Germany passed the Network Enforcement Act in 2018 to better address the dissemination of harmful content online, including incitement to violence. The legislation imposes a 24-hour timeline on social media companies to remove 'manifestly unlawful' content or face heavy fines (up to 50 million euros) if they fail to comply. In the wake of the Christchurch terrorist attacks, Australia swiftly introduced legislation that criminalises "abhorrent violent material". If social media companies fail to remove such content expeditiously, they could face fines of up to 10 per cent of their annual profit, and employees could be sentenced to up to three years in prison.[15]

13 https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html?utm_term=.12054a893c92

14 https://www.dia.govt.nz/Response-to-the-Christchurch-terrorism-attack-video, https://www.netsafe.org.nz/advisory16march2019/, https://www.radionz.co.nz/news/political/386237/current-hate-speech-law-very-narrow-justice-minister-andrew-little, https://www.stuff.co.nz/national/104126249/all-you-need-to-know-about-the-proposed-privacy-laws?rm=a

15 https://www.theguardian.com/media/2019/apr/04/australia-passes-social-media-law-penalising-platforms-for-violent-content

The UK Government is also looking to address harmful content online. Before deciding what action it will take, it has set out 'a rolling programme of work to agree norms and rules for the online world and put them into practice' through its Digital Charter.[16] As part of this process, it undertook a public consultation in 2017 on its Internet Safety Strategy Green Paper[17], and in April 2019 released an Online Harms White Paper for consultation.[18] In the White Paper, the UK Government states that it will establish a new statutory duty of care to make companies take more responsibility for the safety of their users and tackle harm caused by content or activity on their services. The UK is proposing to regulate not only social media companies, but also any company that "provides services or tools that allow, enable, or facilitate users to share or discover user-generated content, or interact with each other online".

Compliance with the new duty of care will be overseen and enforced by an independent regulator which will set out how companies must fulfil their duty of care through codes of practice, and will have a suite of powers to take effective enforcement action against companies that have breached their statutory duty of care, including the powers to issue substantial fines and to impose liability on individual members of senior management. In the most serious of cases, the regulator may be given powers to force third parties to disrupt a non-compliant company's business activities or internet service providers to block access to certain websites. The regulator will also have the power to require annual transparency reports from companies, which outline the prevalence of harmful content on their platforms and what counter measures they are taking to address these, and additional information, including about the impact of algorithms in selecting content for users and to ensure that companies proactively report on both emerging and known harms.

Another key example of efforts to work with social media companies – or take a co-regulation approach - to stem online incitement to violence or hatred against certain groups is the European Commission's (EC) Code of Conduct on Countering Illegal Hate Speech Online (Code of Conduct).[19] The Code of Conduct is a voluntary, non-binding set of commitments designed to combat the spread of illegal hate speech online in Europe. Signatories commit to developing processes to review notifications regarding illegal hate speech on their services so they can remove or disable access to such content within 24 hours from their platforms; educate and raise awareness with their users about the types of content not permitted under their rules and community guidelines; and share best practices with each other. Facebook, Twitter, YouTube and Microsoft have signed up to the Code of Conduct with other social media platforms, including Instagram, Google+ and Snapchat, announcing

16 For more information on the Digital Charter, please see https://www.gov.uk/government/publications/digital-charter

17 For more information on the consultation, please see https://www.gov.uk/government/consultations/internet-safety-strategy-green-paper

18 For more information on the UK Government's Online Harms White Paper, please see https://www.gov.uk/government/consultations/online-harms-white-paper

19 For more information on the European Commission's Code of Conduct on Countering Illegal Hate Speech Online, please see https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300

their intention also to join.[20] In assessing whether this approach is working, according to the fourth monitoring exercise of the implementation of the EC's Code of Conduct social media companies had an 89 per cent success rate in reviewing notifications within 24 hours, a significant increase from when the Code of Conduct was launched in 2016 when it stood at 40 per cent within 24 hours).[21]

For terrorist content online, however, the EC is proposing to take a tougher approach. It is developing new rules that will oblige social media platforms to remove terrorist content or disable access within one hour of receiving a removal order from authorities. If a hosting service provider fails to comply with removal orders, they may be liable to a penalty of up to a maximum of 4 per cent of their global turnover for the previous year.[22] In the aftermath of the Christchurch terrorist attacks, the EC has underscored the importance of adopting the new regulations as soon as possible.[23]

Legislation and regulations that impose tight timelines and heavy sanctions on social media companies, however, have been roundly criticised by freedom of speech advocates. The key concern raised is that in the face of high fines social media companies will take a 'play it safe' approach and will be more likely to delete potentially unlawful or harmful material, which may be legitimate expressions of opinion.[24] Additionally, there are also concerns that such provisions delegate too much power to social media companies to determine what is protected under the right to freedom of expression and what is not, which should be subject to judicial determination and due process.[25]

In the lead up to the release of the UK's Online Harms White Paper, critics raised similar concerns and called on the UK Government to take a 'human rights by design' approach towards any legislation, regulation, or other measures it develops to reduce online harms, in order to ensure that the right balance is struck between respecting human rights and meeting the legitimate interests of governments in having unlawful and harmful content removed.[26] One key proposal in this respect is to create a new model of oversight, which combines industry-developed standards with a multi-stakeholder mechanism for enforcement – called an Independent Online Platform Standards Oversight Body – to provide transparency, accountability and representation of the public interest.[27]

20 As above.

21 European Commission 2019, Code of Conduct on Countering Illegal Hate Speech Online – Results of the fourth monitoring exercise, available at: https://ec.europa.eu/info/sites/info/files/code_of_conduct_factsheet_7_web.pdf

22 For more information on the European Commission's Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online, please see https://www.consilium.europa.eu/en/press/press-releases/2018/12/06/terrorist-content-online-council-adopts-negotiating-position-on-new-rules-to-prevent-dissemination/

23 https://www.euractiv.com/section/digital/news/eu-institutions-at-loggerheads-over-online-terrorist-content/

24 https://www.gp-digital.org/news/gpd-provides-briefing-to-uk-government-on-upcoming-online-harms-white-paper/

25 https://www.eff.org/deeplinks/2019/02/eus-proposal-curb-dissemination-terrorist-content-will-have-chilling-effect-speech

26 https://www.gp-digital.org/news/adopt-a-human-rights-by-design-approach-towards-regulating-online-content-say-civil-society-groups/

27 Global Partners Digital 2018, A Rights-Respecting Model of Online Content Regulation by Platforms, available at: https://www.gp-digital.org/publication/a-rights-respecting-model-of-online-content-regulation-by-platforms/

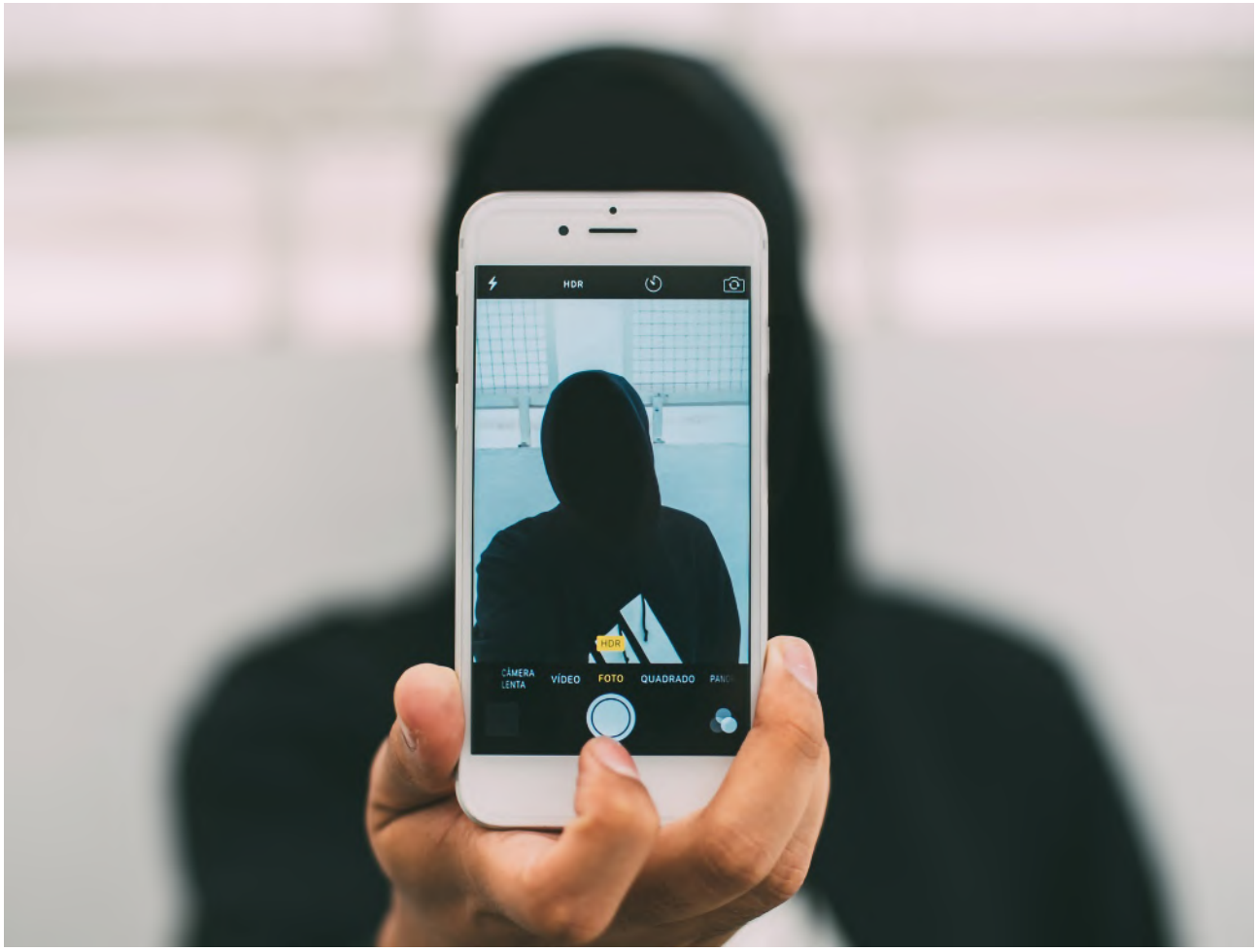


Image credit: Kaique Rocha.

As now recognised in the UK Government's Online Harms White Paper and Australian legislation, there are growing calls for social media platforms to be subject to a statutory duty of care to give better protection against online harm. If subject to a statutory duty of care, social media companies would be obliged to take reasonable care to ensure that their users are reasonably safe when using their platform, and that they prevent, or reduce the risk of, one user harming another. To these ends, social media companies would need to invest in and take reasonable measures to prevent harm by, for example, improving their technology-based responses and/or by making changes to their terms of service, otherwise they would face penalties from a regulatory body mandated to oversee and monitor online harms.[28]

28 https://www.carnegieuktrust.org.uk/blog/social-media-regulation/

New Zealand's current policy and legal framework for addressing terrorist and harmful content online

Social media companies in New Zealand are largely left to self-regulate how they monitor and remove harmful content in line with internal policies and guidelines. If the content is objectionable, a patchwork of New Zealand agencies has limited oversight powers – these agencies include the Privacy Commission, the Ministry of Justice, the Department of Internal Affairs and Netsafe. As we saw in relation to the Christchurch terrorist attack, the first line of defence is when social media companies are prompted to remove terrorist and harmful content from their platforms by receiving reports from users or authorities or the content triggers their internal monitoring systems.

If this line of defence fails, while there are numerous legislative measures already in place to deter people from distributing harmful content online, there are limited provisions specific to social media companies.

As we have seen in the aftermath of the Christchurch terrorist attack, the video of the attack and the alleged perpetrator's manifesto have now been classified as objectionable content by the Chief Censor under the Films, Video and Publication Classification Act 1993, and anyone who views, possesses or distributes either could face criminal charges. It is unclear if the Act applies to social media companies, or whether they are protected by its exemptions relating to distribution.[29] The information provided on the Department of Internal Affairs' (DIA) website – that it has informed social media platforms that distributing the video, which includes hosting it, is an offence – suggests they are not exempt. So far, however, DIA's approach has been to ask social media platforms to remove it voluntarily. It seems they have 'responded positively' as it is objectionable and does not align with their terms of service.[30]

Similar charges could also be brought under the Harmful Digital Communications Act 2015, which makes it an offence for anyone to cause harm intentionally by posting a digital communication which leads to 'serious emotional distress' (noting that the main purpose of this Act is to protect against online harms such as cyberbullying, revenge porn, and other forms of harassment and intimidation rather than against online hate speech). Offenders can face up to two years in imprisonment or a fine up to $50,000.

Social media companies as 'online content hosts' are protected from liability under the Act, however, provided that they follow the 'safe harbour' process for handling complaints (including actioning a complaint within 48 hours of receipt).[31] The Act explicitly states that failing to comply with this process does not create a liability on social media companies. Whether the safe harbour provisions of the Harmful Digital Communications Act 2015 need to be

29 See section 122 of the Film, Videos and Publications Classification Act 1993.

30 Department of Internal Affairs, The Department's Response to the Christchurch Terrorism Attack Video – Background information and FAQs, available at: https://www.dia.govt.nz/Response-to-the-Christchurch-terrorism-attack-video#ISPs

31 For more information on the safe harbour process, please see https://www.justice.govt.nz/justice-sector-policy/key-initiatives/harmful-digital-communications/safe-harbour-provisions/

strengthened to increase the liability of social media companies and impose shorter removal timelines and sanctions for non-compliance – like the German and EC approaches discussed above – warrants further consideration.

Under the 'racial disharmony' provisions in New Zealand's Human Rights Act 1993, it is a civil offence to publish or distribute written material or broadcast by electronic communication words which are threatening, abusive, or insulting and are likely to excite hostility against or bring into contempt any group of persons in New Zealand on the ground of colour, race, or ethnic or national origins. If a person intends to incite racial disharmony, a criminal prosecution can be brought, but only with the consent of the Attorney-General. These provisions would not apply to the alleged perpetrator of the Christchurch terrorist attack, however, as religion is not recognised as grounds for prosecution. In the absence of a hate speech prosecution, the alternative would be to charge the alleged perpetrator with a hate crime. New Zealand law, however, does not specifically provide for hate crimes motivated by religion or on any other grounds, although such grounds can be considered as an aggravating factor under the Sentencing Act 2002. Whether New Zealand law should specifically recognise hate crimes has been raised again in the wake of the Christchurch terrorist attack, as has the need for police to monitor better whether crimes were hate-motivated to understand the extent of the issue fully.[32]

The Minister for Justice, Andrew Little, has acknowledged the need to urgently review the Human Rights Act 1993, as well as New Zealand's lack of hate crimes laws, in response to the Christchurch terrorist attacks.[33] In the meantime, only the Harmful Digital Communications Act 2015 - which recognises that "a digital communication should not denigrate an individual by reason of his or her colour, race, ethnic or national origins, religion, gender, sexual orientation, or disability" as one of its Communication Principles – specifies that those with decision-making powers under the Act, including the court, must take these grounds into account when determining whether a person has committed an offence under the Act.

Reform current laws or take a bolder approach? The need to intervene and regulate social media companies in New Zealand

As outlined above, there are key gaps in New Zealand's current legislation with respect to addressing harmful content online and holding social media companies to account for its distribution. There is room to amend and strengthen existing legislation – such as the Films, Video and Publication Classification Act 1993 and the Harmful Digital Communications Act 2015 – to increase the liability of social media companies and impose higher financial penalties for non-compliance. Considering the extent to which social media companies are left to self-

32 https://www.radionz.co.nz/news/national/386102/law-change-should-consider-protecting-religious-groups-against-hate-speech

33 https://www.stuff.co.nz/national/christchurch-shooting/111661809/hate-crime-law-review-fasttracked-following-christchurch-mosque-shootings

regulate in New Zealand and that it is not in their commercial interests to restrict the spread of harmful content online, however, we recommend that the New Zealand Government takes a comprehensive regulatory response to ensure that social media companies are appropriately incentivised to act in the public interest and address harmful content online effectively.

We recommend that the New Zealand Government consider following a similar path to that of the UK and establish a new regulatory body to oversee social media companies operating in New Zealand, which can set standards governing the distribution of harmful content online and ensure social media companies abide by them.

Depending on the level of regulatory intervention favoured by the Government, it could call for New Zealand's social media industry to establish its own independent regulatory body, like the New Zealand Media Council (NZMC), to set industry standards and resolve complaints. Like the NZMC model, its success would depend on the voluntary co-operation and compliance of its member organisations (as it would have no statutory powers to enforce its decisions or impose sanctions) and would be self-funded by its members.

In the present context, the Government may favour a co-regulatory body, like the Broadcasting Standards Authority (BSA), which is established by statute and all social media companies in New Zealand are subject to its jurisdiction. Like the Broadcasting Act 1989, legislation could be drafted that contains a set of common standards that all social media companies in New Zealand must abide by and provides powers to the oversight body to adjudicate complaints on alleged breaches of the standards and determine penalties. Like the BSA, the oversight body's functions could also include issuing advisory opinions, working with the social media industry to develop codes of practice, and conducting research on matters relating to standards in social media. The oversight body could be funded by an industry levy in the same way the BSA is funded. Another consideration is whether the oversight body should be given powers to initiate investigations into significant breaches of standards by social media companies rather than only if a complaint from a member of the public is received (as is the current limit on the powers of both the NZMC and the BSA).[34]

As part of these reforms, we recommend the New Zealand Government also imposes a statutory duty of care on social media companies operating in New Zealand. In the wake of the Christchurch terrorist attack, there have been calls for social media companies to "fulfil their duty of care to prevent harm to their users and to society".[35] As New Zealand's Privacy Commissioner has stated, sharing graphic content is a predictable risk of a livestreaming feature.[36] As is currently being proposed in the UK, the New Zealand Government could consider imposing a statutory duty of care on social media companies to ensure they take more responsibility for the safety of their users

34 For a more detailed analysis of the NZMC and the BSA governance models, please see New Zealand Law Commission 2013, The News Media Meets New Media, available at: http://r128.publications.lawcom.govt.nz/

35 https://www.newsroom.co.nz/2019/03/20/498595/govt-fund-managers-call-for-social-media-changes-after-fridays-massacre

36 https://www.radionz.co.nz/news/national/385072/christchurch-mosque-attacks-predictable-risk-in-facebook-livestreaming-feature

and tackle harm caused by content or activity on their services. In terms of penalties, considering that social media companies are driven by commercial imperatives, we recommend that penalties are set at a level that will incentivise social media companies to combat harmful content online effectively. In line with the approach taken in Australia and being considered by the European Union and the UK, the New Zealand Government should consider imposing penalties which oblige social media companies to forfeit a percentage of their annual revenue. At the same, as noted above, appropriate checks and balances must also be put in place to ensure the right to freedom of expression is not impacted by social media companies being subject to high penalties for non-compliance. This could be achieved by having a wide variety of interests represented on any enforcement body.

As part of this programme of work, we also recommend that the New Zealand Government carefully considers the extent to which the current law protects against hate speech. As a matter of principle, the New Zealand Government should amend the Human Rights Act 1993 to ensure that its hate speech provisions also apply to religion, and other grounds of discrimination contemplated by the Act. The extent to which hate-motivated crimes occur in New Zealand and whether they are appropriately provided for in legislation also warrants detailed review.

Image credit: Kristina Hoeppner.

MITIGATING TERRORIST AND HARMFUL CONTENT ONLINE: NEED TO EFFECTIVELY ADDRESS HATE SPEECH AT THE SAME TIME

While finding better ways to prevent the upload and spread of terrorist and harmful content online must be a top priority, careful consideration also needs to be given to what earlier interventions could be taken to avert a similar event, like the Christchurch terrorist attack, from occurring in the future.

The potential role that the internet plays in providing space for people to share extremist views and how that leads to the real-life violence against others and terrorist acts must not be overlooked. While it is difficult to establish a causal link between online hate speech and violence against minorities or terrorist acts, research suggests that in the case of terrorist content and radicalisation, sustained exposure may reinforce beliefs that are already extreme, and violent, hateful language can inflame people who are already inclined toward violence and focus their rage.[37] Online hate speech can also influence people in a similar way. A recent study in Germany found that an increase in anti-refugee rhetoric on Facebook was correlated with an increase in physical attacks against refugees.[38] The researchers concluded with the suggestion that "social media has not only become a fertile soil for the spread of hateful ideas but also motivates real-life action." The connection between online hate speech and violence against minorities has also been thrown into stark relief in Myanmar, with the UN Independent Fact-Finding Mission on Myanmar finding that social media, in particular Facebook, had played a 'determining role' in the human rights violations committed against the Rohingya population in spreading disinformation and hate speech. [39]

In addition to stamping out illegal hate speech online, there is also a need to better monitor hate speech activity on social media as a key early intervention mechanism. It has been reported that the alleged perpetrator of the Christchurch terrorist attack spread his manifesto widely online and was active on social media in the days leading up to the attacks. He also forewarned what he was about to do on Twitter and 8chan (an anonymous forum known for its politically extreme and often hateful commentary) before livestreaming his attack.[40] It is likely that some of these users were responsible for the rapid dissemination of the video on social media platforms and altering its format to avoid detection.

It is unclear to what extent New Zealand's intelligence services were monitoring white supremacist or right-wing extremist groups in the lead up to the Christchurch terrorist attack. Both the Security Intelligence Service (SIS) and the Government Communications Security Bureau (GCSB) confirmed the alleged perpetrator was not on their radar and they had not collected or received relevant intelligence in the lead up to

37 https://www.rand.org/randeurope/research/projects/internet-and-radicalisation.html

38 Müller, K. & Schwarz, C. 2018, Fanning the Flames of Hate: Social Media and Hate Crime, available at: https://ssrn.com/abstract=3082972

39 https://www.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUSKCN1GO2PN

40 https://www.washingtonpost.com/technology/2019/03/15/facebook-youtube-twitter-amplified-video-christchurch-mosque-shooting/?noredirect=on&utm_term=.1b90a539afac

the attacks. An analysis of 10 years of public documents from the SIS and the GCSB also found that the threat posed by white supremacists or right-wing extremists was never specifically mentioned.[41] This apparent lack of attention to far-right extremism has contributed to public distrust in the agencies and must be addressed. The extent to which New Zealand's intelligence services had the adequate powers, capacity and funding to detect this type of activity in the lead up to the Christchurch terrorist attack, as well as whether they accurately categorised and prioritised these as security threats, requires closer scrutiny and will be a key focus of the Royal Commission. The role that social media companies can play in better monitoring and removing hate speech online – keeping in mind the freedom of speech considerations noted above – and working more effectively with intelligence services to avert extreme harm online also needs to be factored in to any measures designed to regulate their practices better in the future.

Image credit: James Dann.

41 https://www.radionz.co.nz/news/political/385173/no-mention-of-right-wing-extremist-threats-in-10-years-of-gcsb-and-sis-public-docs?utm_source=The+Bulletin&utm_campaign=d4d7f92d23-EMAIL_CAMPAIGN_2018_03_01_COPY_01&utm_medium=email&utm_term=0_552336e15a-d4d7f92d23-533756713

NEED FOR A GLOBAL RESPONSE TO TERRORIST AND HARMFUL CONTENT ONLINE

While addressing terrorist and harmful content online has become a key discussion point in New Zealand after the livestreaming of the Christchurch terrorist attack, it is also a global issue that requires an internationally-co-ordinated response. In the wake of the Christchurch terrorist attack, New Zealand's Prime Minister Jacinda Ardern is spearheading efforts to build a global coalition to address terrorist content online. Alongside the 'Tech for Humanity' meeting of G7 Digital Ministers in Paris on 15 May 2019, Prime Minister Ardern and French President Emmanuel Macron will host a summit to rally technology companies and other concerned countries to commit to a pledge – referred to as the 'Christchurch Call' – to eliminate terrorist and violent extremist content online.

As outlined above, several countries and the European Union have already developed or are working on a variety of policy responses to combat the spread of terrorist and harmful content online. It is crucial that the 'Christchurch Call' consolidates these approaches and leads to an international plan of action that provides a clear and consistent framework on the issue, including what constitutes terrorist content, how freedom of speech will be safeguarded, how quickly platforms should be forced to take down illegal content and what penalties they will face for non-compliance. The framework should also complement and link to other global initiatives to counter terrorism online, including the UN's Plan of Action to Prevent Violent Extremism which outlines key actions in relation to the internet and social media, the Tech Against Terrorism initiative, and the industry-led Global Internet Forum to Counter Terrorism. Considering the importance of combatting hate speech as a means of preventing terrorist and extremist content online, the 'Christchurch Call' should also acknowledge and prioritise the implementation of the Rabat Plan of Action, which outlines member states' international human rights obligation to prohibit "advocacy of national, racial or religious hatred that constitutes incitement to violence, hostility or discrimination".

Some of the biggest technology companies, including Google, Facebook, Microsoft, and Twitter, are expected to participate in the Summit, which presents a key opportunity to develop a multi-stakeholder approach to the issue. As part of the 'Christchurch Call', countries should also seek clear commitments from technology companies on what practical actions they will take to find new and innovative ways to catch terrorist and harmful content that is currently difficult to detect, like the Christchurch terrorist attack video, and minimise its distribution.

Image credit: CHRISTCHURCH CITY COUNCIL.

Q AND A WITH PRIVACY COMMISSIONER JOHN EDWARDS

Is the current patchwork of social media regulation in NZ working?

No. The patchwork means that there is an inconsistency of application of New Zealand domestic legislation to agencies providing online services to, but not based in, New Zealand, and an inconsistency of regulatory or enforcement mechanisms to apply even if jurisdiction is clearly established.

This might be because no one framework is specifically focused on the medium (i.e. the platforms). Rather they apply to different types of content and each has a slightly different objective and focus. The Films, Videos, and Publications Classifications Act applies to certain specified types of prohibited harmful content on the basis that they might cause societal harm. The Human Rights Act is focused on protecting specified groups from certain specified harms. The Privacy Act is about the use of personal information and provides remedies to individuals who are harmed for misuse of their personal information, but does not provide for effective regulation of conduct using personal information that might undermine social institutions, or foment social discord in the absence of a specific harm being experienced by an individual. The Harmful Digital Communications Act seeks to provide an ability to individuals to have harmful content removed when in it directly affects an individual, but protects the platforms through the safe harbour provisions, wherever it is held or distributed.

This fragmentation allows social media companies to avoid responsibility for the content. They believe it is appropriate to avoid responsibility both as a matter of principle (CDA 230, and the First Amendment) and because of the difficulties of scale. The first objection is a claim that US-based values should define online conduct in all jurisdictions, and that domestic lawmakers should have no role and be passive "takers" of both technology and the terms under which it is offered.

The second is sophistry and a self-serving response to a problem entirely of the successful platform's own making. It implies that it is reasonable to saddle a small emerging (and potentially disruptive) online business with the burden of responsibility for content, but large established services should be exempt because it is too difficult for a big company to comply with legal norms of the jurisdictions in which they operate. In no other industry would such a preposterous proposition be entertained.

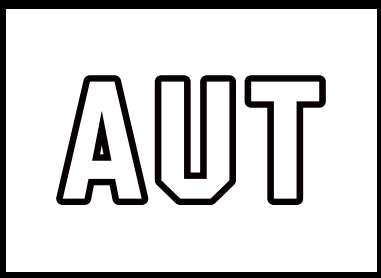
What changes are needed in your view to prevent this kind of event happening again?

A domestic response needs to be comprehensive and addressed at the platforms, holding them to account for their risk taking (such as rushing a new product like live streaming to market without adequately mitigating obvious risks). However this needs to be coupled with an international response to ensure mutual recognition and enforcement of rights where the actors refuse to accept the legitimacy of any domestic legislation.

Is a statutory duty of care (as has been proposed in the UK) something you believe has potential in NZ?

Yes. Individuals and classes should have the right to sue where the duty of care has been breached. A regulator should be able to take action where no individual or class has standing or motivation, but the breach has the potential to cause public harms. For example, the Cambridge Analytica scandal led to a large number of individuals being influenced by personalised and targeted advertising to "nudge" their behaviour. If a neo-Nazi gets a message to support a particular candidate, or an apathetic and sceptical voter of a likely political hue gets a message telling them not to bother showing up to vote because it will be pointless anyway, neither is likely to make a credible claim for experiencing harm from that manipulation, or be motivated to do so, notwithstanding that that abuse of the platform and the data might cause serious societal harm and the undermining of democratic institutions.


THE
Helen
Clark
FOUNDATION


AUT