AUGUST 2021

REPORT TO THE INVESTOR COLLABORATION THAT HAS BEEN ENGAGING WITH SOCIAL MEDIA COMPANIES IN RESPONSE TO THE CHRISTCHURCH TERROR ATTACKS:

Assessing content moderation during objectionable content crises by Facebook, Twitter, Alphabet, and the trajectory of regulation

BRAINBOX

About Brainbox

Brainbox is an independent consultancy and think tank based in New Zealand, which specialises in issues at the intersection of technology, politics, law and policy. Brainbox's investigations are available at www.brainbox.institute and cover matters such as legislation as code, judgments as data, trust and automated decision-making, and the legal implications of synthetic media.

Editing support and layout were provided by **Antistatic** (antistaticpartners.com).

Disclaimer

This report was commissioned by the Guardians of New Zealand Superannuation, with support from Neuberger Berman, and Northern Trust. The contents of this report are the outcome of Brainbox research and analysis, and do not necessarily reflect the views of the funders or other individuals referenced or acknowledged within the document.

Brainbox's brief was to apply its expertise in this and related subjects in order to reach conclusions and provide key insights to the group on the two key questions identified in parts 1 and 2. Brainbox's role has not been to provide advice or recommendations to the investor group, whose members each have their own priorities and obligations when it comes to responsible investment practices and their relationships with the platform companies. Brainbox has formed its opinion based on desktop research, academic papers and publicly available resources.

Recommended citation

Brainbox (2021). Report to the investor collaboration that has been engaging with social media companies in response to the Christchurch terror attacks.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

CONTEXT

The perpetrator of the 15 March 2019 Christchurch terror attack filmed and “live-streamed” (broadcast via the internet) his attack. Subsequent videos were copied, modified, and widely disseminated online. The videos significantly increased the scale of the terrorist effect.

Significant work has taken place with the aim of preventing the spread of this kind of violent objectionable content online in connection with a real world terror attack. In this investigation, our focus was on the online components related to the terror attack, not the terror attack itself. We refer to the kind of widespread dissemination of objectionable content such as that which related to the terror attacks of 15 March 2019 as an **objectionable content crisis (OCC).**

This report deals with two inter-related but distinct topics, which we address in two parts:

1) Part 1: in light of the terror attacks in Christchurch of 15 March 2019, are the changes made by Facebook, Google/YouTube (Alphabet), and Twitter sufficient to prevent or mitigate the risk of similar objectionable content being created, accessed, and shared at similar scale? Part 1 has been carefully scoped to narrow our investigation and these scoping decision are explained in an appendix.
2) Part 2: Nation States are increasingly proposing that the activities of digital platform companies such as Facebook, Twitter and Alphabet should be regulated. Some of this is a direct response to the 15 March terror attack. What is the broad direction of regulatory travel, what are some pros and cons of the various proposals, and what does good regulation look like?

Part 1 primarily relates to the changes implemented in response to the online proliferation of audiovisual content that formed part of the 15 March terror attack. It also considers the platforms’ relationships with users and other platforms. Part 2 considers how nation states are using the law to regulate the way that platforms interact with their users.

PART 1: PLATFORM CHANGES

In Part 1, we conclude that the platforms have adopted a range of collaborative measures which are likely to be very effective at mitigating or preventing future objectionable content crises (OCC) of the kind that occurred on 15 March 2019. However, these measures are unlikely to entirely prevent all future OCCs.

All the measures introduced have trade-offs and limitations. It will be a matter of ongoing refinement as to how these are to be balanced against efforts to improve the efficacy of future OCC responses. The best way of ensuring that this balancing process supports the public interest will be to invite independent scrutiny and assessment of how the measures are being implemented. By way of summary and conclusion we note:

- Multi-platform collaborative measures play the greatest role in enabling platforms to rapidly classify and intervene in new objectionable content during an OCC. Conversely, platform collaboration presents risks to human rights and requires measures to enhance auditability and transparency of such collaborations.

- The most effective measures of limiting an OCC are: the Global Internet Forum to Counter Terrorism (GIFCT) Content Incident Protocol (CIP) and the use of the GIFCT shared hash database. The greatest limitations of the CIP and the shared hash database are the ways that they use automation to remove content rapidly, creating risks of unjustifiable and inscrutable content removals. Further, there is a risk that these mechanisms could be abused by the platforms or by the influence of nation states. While transparency measures are an important safeguard, transparency is difficult to achieve these systems must maintain some secrecy in order to avoid gaming or abuse by perpetrators.
- In situations where a crisis falls short of the activation requirements for a CIP, the two crisis response protocols developed by GIFCT and tested through tabletop exercises with government and civil society will play an important role. It is difficult to assess the effectiveness of these response protocols from an external perspective.
- Platforms must remain focused on enhancing the speed with which they can reliably detect and classify content.[1] For this reason, ongoing improvements in content moderation systems are an essential prerequisite for responding to an OCC. Importantly, the speed with which these systems can classify content should not come at excessive cost to the accuracy and reliability of these systems. Investors could support the improvement of content moderation systems by advocating for measures which enhance transparency and support the development of shared bodies of expertise in how content moderation is conducted at scale.

We are not persuaded that any changes by the platforms to decrease the accessibility of livestreaming services will have a significant impact in limiting future OCCs. We also note that the trade-offs of limiting public access to this technology are significant.

The platforms are all constantly engaged in a range of improvements to their content moderation systems and procedures. These improvements are fundamental not just to the platforms' ability to respond to crises, but also to their basic viability as platform businesses for receiving and delivering user-generated content. It is impossible to accurately catalogue and assess each of these across the three companies from an external perspective. This is one reason why we endorse regulatory approaches which standardise and formalise transparency and auditing metrics around platform content moderation. This would produce reporting data that can support external analyses of the kind we have undertaken here.

While the platforms have made a range of changes to respond to the 15 March terror attack OCC and similar events, we observed frequent calls for platforms to be more transparent about the methods they have adopted. In particular, there were frequent calls to enhance independent researcher and auditor access to key institutions and datasets, including the platforms themselves and the GIFCT. We note that in July 2021, GIFCT released a human rights impact assessment of itself as well as announcing a range of new initiatives. We have not examined these in detail, but they appear to provide a useful platform for future improvements.

[1] We use "classify" in the sense adopted by Gorwa et al as being related to assigning content into particular categories for moderation purposes. See Gorwa R, Binns R and Katzenbach C, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945.

PART 2: REGULATORY TRAJECTORY

In Part 2, we identify the current trajectory of regulatory controls on online content moderation — that is, the aggregate direction of current and emerging content moderation regulation (specifically legislation). We also carry out an analysis of this trajectory. Our conclusion is that the strongest case for regulation requires transparency and auditability of content moderation systems. We cannot endorse content-specific standards in regulation, unless these are linked to content which is already illegal (for example, child sexual abuse material or an incitement to violence). We also urge caution about heavy handed or punitive approaches imposed on platforms that incentivise content moderation practices that are inconsistent with human rights, unjustifiably rely on automation, or fail to consider perverse incentives.

The trajectory we identified for the development of online content moderation regulation is:

States are increasingly looking to regulate social media content

1. The activity of the platforms is beginning to touch upon the interests of nation states. It is also beginning to affect the rights and interests of citizens within those states' sovereign jurisdictions. These states therefore have a legitimate interest in regulating the platforms, insofar as all states can justifiably limit some human rights and interests in order to protect other human rights and interests. A core constraint here is that states may only limit human rights in a manner that itself complies with human rights norms and principles. We expand upon this below.
2. The predominant trend in regulation is toward the use of legislation (ie use of law) to control how the platforms moderate content. It is difficult to extricate regulation that affects content moderation from other areas of law and policy, such as antitrust, privacy, the use of AI systems, "honest advertising", election interference, misinformation, and other areas.

Human rights create the appropriate framework for saying what "good" and "bad" looks like in this emerging area

3. Content moderation is an emerging area. As such, there are few established assessment standards. While the human right to freedom of expression is an old topic, the introduction of digital platforms raises many new issues. As a result, consensus is still emerging on two points:
    a. How should the platforms be moderating particular kinds of content, particularly in a global context? This question has a procedural element as well as a substantive element. Specifically, it asks what kinds of content should be impermissible, but it also asks what procedures should be followed by platforms and by states in moderating that content. This makes regulating difficult because it is difficult to clearly and specifically identify and then say what we want platforms to do and how they must do it.
    b. What is the proper role of government when it comes to using law to influence how the platforms moderate content produced by users? Many proposed laws set a role for governments in directing the platforms to moderate content in a particular way. Because the appropriate role of a government in this situation is not clear, it is difficult to say whether these proposed laws allocate appropriate rights and responsibilities to platforms, states and users.

4. Our research indicates that there is widespread support from across the spectrum of stakeholder groups for turning to human rights instruments, principles and jurisprudence to generate greater consensus on the questions we outline above. The UN Declaration of Human Rights (and various associated instruments) outlines a universal set of standards which are intended to manage the relationships between the rights of individuals and states (and increasingly, commercial entities). Human rights instruments set out a widely agreed statement about what can and should be done by States when it comes to balancing the rights of individuals, including both users of platforms and the platforms themselves. There should be ready agreement that regulation which undermines human rights without justification is undesirable. Equally, regulation which requires the companies themselves to undermine human rights is also undesirable.

Human rights can be justifiably limited and balanced, but only according to human rights principles

5. Human rights jurisprudence sets out ways for states to justifiably limit human rights in order to protect other interests. In summary, to limit the human rights of individuals using law, states must comply with the principles of *legality*, *legitimacy*, *necessity* and *proportionality*.
6. Much of the enacted and prospective regulation has the potential to limit a range of human rights. For this to be justifiable, the regulation must be in accordance with human rights principles in the following ways:
    a. Pursuant to the principles of necessity, proportionality and legitimacy, there must be a demonstrable connection between a proposed regulatory intervention and an adverse outcome to a legitimate interest protected by human rights instruments. These adverse outcomes should be real (not hypothetical) and significant enough to outweigh the harms caused by limiting a human right. This means that states must be able to: persuasively show that the thing they are seeking to limit is causing a real adverse outcome; of the kind the state can legitimately protect against under human rights law; that state intervention is necessary to avoid the adverse outcome; that state intervention will in fact mitigate or avoid the adverse outcome; and that there are no less invasive methods available to achieve the same effect. To put it bluntly, States have to show that particular content is truly undermining peoples' human rights - it is not enough to point to a vague connection between content and an alleged or hypothetical harm. For some types of content this will be easy, but for others, it will not.

    b. Pursuant to the principle of legality, it is extremely difficult to articulate clear and reasonably unambiguous categories of content. This means that, even if States have a clear idea of the kind of content they are targeting, it will be difficult to use language to articulate that category in a predictable way. Furthermore, even in a best case scenario, the assessment of whether content falls into a category will involve complex matters of fact and law that must be resolved on a case-by-case basis, require time and resourcing by the platforms to apply correctly, and still presents significant risk of error. Even where these categories can be applied correctly, the principle of legality still requires procedural rights of review and appeal to legal bodies. As such, the potential volume of legal cases generated by compliance with a regulatory regime may be enormous.

7. In some cases, a human rights approach means that states should not intervene to prevent harm. In other words, a person might harm another person (e.g. by an action or utterance), but preventing this would be a breach of human rights by the state. This proceeds from the starting point that the general ability to act, speak, and think is essential to human dignity, and therefore protected by human rights instruments except in narrow circumstances. A human rights approach is concerned with the appropriate balancing of various adverse outcomes, not the total avoidance of harm. If states wish to justify regulatory interventions, then the best thing they can do is to support empirical work exploring the connection between particular types of content and real adverse outcomes. This is necessary to perform the balancing exercise which is core to human rights approaches.

8. Some of the regulatory proposals we examined would require the platforms to systematically breach the human rights of their users, for example by creating compliance conditions that are so strict and punitive that platforms are simply unable to moderate content according to human rights principles. This often proceeds from regulators and politicians holding unjustified confidence in the capabilities of automated content moderation systems. There is widespread opposition from a range of groups across the stakeholder spectrum towards most of the regulatory proposals we examined, particularly toward Australia's Abhorrent Violent Material amendments and the EU "24 hour" terrorist content proposal. We believe this opposition is justified. We also found cautious and appropriate support from human rights bodies for other regulatory proposals, particularly the EU Digital Services Act.

POTENTIAL NEGATIVE OUTCOMES

If states pursue regulation of online content according to the worst features of the current trajectory — that is by using legislation to prohibit broad categories of expression which is not illegal, or by imposing unrealistic compliance conditions — then we anticipate the following negative outcomes:

1. More unjustified automated takedowns of content, including higher rates of false positive takedowns by algorithmic systems that embed human social biases.[2]

2. More frequent objections by human rights bodies.

3. Extensive litigation between platforms and users, between platforms and governments, and between users and governments.in a range of different courts and tribunals globally.

4. Continued public outcry against platform content moderation, but on different points. If transparency obligations are not enhanced and implemented, then much of this public outcry will remain situated around anecdotal cases, fuelling outrage without offering meaningful opportunities for insights and progress.

5. A potential exodus of users, whether away from the platforms entirely, toward other platforms, or toward the encrypted applications offered by these platforms and others.

[2] Shenkman, C., Thakur, D., Llansó, E. (2021) Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis. Center for Democracy & Technology. <https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/>.

6. Handing powers to states that leads to abuse in particular cases, or more widespread abuse generally, as the relationship between platforms and governments becomes tighter and the threat of sanctions greater.
7. The global legitimisation of regulatory regimes that empower states to use digital infrastructure to suppress free expression, free association, and free thought, and to identify individuals. This will result in authoritarian states increasing their uptake of such regulatory regimes. By adopting similar legislation, democratic states will be taking a position that weakens their ability to oppose this behaviour in other states. This is because they can be justifiably accused of adopting that behaviour themselves by limiting the human rights of their citizens.
8. An increased pressure on platforms to moderate their software and processes according to local conditions, with the risk that they become extensions of state power with the digital tools to suppress human rights of expression and association at digital scale.

THE ROLE OF NON-STATE REGULATORY APPROACHES

Historically, states have been the greatest threats to the human rights of individuals. This is still the case globally, including in liberal democracies, and despite the widespread acknowledgement that businesses can also influence human rights conditions. For this reason, taking a human rights approach to platform content moderation will frequently lead to the conclusion that states must restrain themselves from passing regulation that threatens human rights. In many cases, for the platforms to really take a human rights approach, they may be required to resist state action, insist upon proof that content is undermining other human rights interests, and to insist on legal and procedural rights that protect platforms and users and restrain the power of the state.

Because states pose a threat to the human rights of privacy, freedom of expression and freedom of association (as well as a range of other human rights), it is worth considering the merits of non-state regulatory options, as well as non-legislative regulatory options. These non-state options do not have to be purely self-regulatory. Regulation can also include the kinds of coordinated interventions we identified in Part 1 of our research (ie, the GIFCT CIP and the shared hash database). If considered through this lens, many “regulatory” interventions have already occurred independently of state action. From a human rights protection approach, it may be preferable to deal with the potential harms of inadequate content moderation through non-state regulatory responses, rather than regulatory responses led by states. It is worth giving some of these existing non-state regulatory mechanisms time to operate, especially because they may also illustrate what works well and does not work well when it comes to state regulation.

POTENTIAL FURTHER ACTIONS

The issue of platform content moderation will become one of the defining legal and socio-political issues of our time, if it has not already. It deserves, and is building, its own body of specialist expertise which stretches across a range of academic disciplines and governance areas. Content moderation is effectively a new label for an extremely old political debate which goes to the core of the political relationship between the individual and the state, which is: what are people allowed to say or think, what can they read or watch, and who gets to decide? This makes it a fraught and volatile issue prone to rash reactions and uncomfortable complexities.

Against this context, there are two means of generating consensus among people and groups who disagree. Our conclusions are that people wishing to see platform content moderation develop in ways conducive to the public interest should pursue the following:

1) Encourage and support empirical investigations that build a better evidential foundation for describing how content moderation is currently conducted, what effects it produces, and how it might be improved. This will include greater access to and transparency by platforms with respect to the roles of independent researchers and audit bodies.
2) In assessing what should or should not be done, adhere to tried and tested legal and political frameworks such as those created by human rights instruments and jurisprudence. These instruments are the product of long and arduous processes of consensus building across a range of social and political groups, and they contain a range of shared foundational insights that can help to shape future discussion.

ENDORSED REGULATORY APPROACH

Based on our limited investigations, and applying human rights principles, our conclusion is that the best kind of regulatory approach is to implement measures that enhance the transparency and auditability of platform content moderation systems and processes. When directed toward transparency and auditability, regulatory approaches of this kind are sometimes accompanied by procedural rights of appeal and review for users against platforms' content moderation decisions. The Digital Service Act proposal advanced by the European Union appears to do the best job of adopting such features, although it is not clear yet how these will be implemented.

We do not endorse regulation that dictates to the platforms what content should be prohibited or taken down (beyond existing legal restrictions against illegal expression such as, without limitation, child sexual abuse material, and incitement to violence).

If states were to introduce regulation that standardised transparency reporting around platform content moderation, and which opened up these systems to independent scrutiny, it would have far-reaching effects. For example, it would provide independent groups, such as human rights and civil society organisations, with tools to ensure that content moderation is being conducted with a view to the public interest and in accordance with human rights principles. It would have the effect of bringing platform content moderation further toward public influence and oversight, while mitigating the risk that the platforms become digital infrastructures for enhancing state control over users' human rights and freedoms.

PART 1: CHANGES MADE BY THE PLATFORMS IN RESPONSE TO THE CHRISTCHURCH TERROR ATTACK

Introduction

The perpetrator of the 15 March 2019 Christchurch terror attack filmed and “live-streamed” (broadcast via the internet) his attack. Subsequent videos were copied, modified, and widely disseminated online. The videos significantly increased the scale of the terrorist effect.

We refer to this kind of widespread dissemination of objectionable content as an **objectionable content crisis (OCC).**

In response, there has been a significant effort to prevent this kind of spread of violent objectionable content online from occurring again, although we note that some initiatives pre-date the attacks.

In Part 1, our brief is to investigate the changes made by Twitter, Facebook and Alphabet (together, “the platforms”) following the 15 March 2019 OCC. The core inquiry is whether these changes are sufficient to prevent or mitigate the risk of similar objectionable content being created, accessed, and shared at similar scale in the future.

Changes made by the platforms to mitigate future OCC are the focus of Part 1 of this report, by contrast with the wider area of “content moderation”, which is investigated in Part 2.

In relation to Part 1, the following scoping decisions have been agreed with the investor group (we explain them in more detail in an appendix to this report):

- Our analysis does not examine the question of how the platforms contributed to the terrorist’s radicalisation, his radicalisation to violence, or the real world terror attacks.
- Our analysis excludes matters relating to the terrorist’s manifesto.
- Our analysis excludes other social media platforms such as Reddit, 4Chan, 8Chan. It also excludes the file hosting websites hosting the manifesto.
- Our analysis in Part 1 excludes government agencies and the topic of government regulation (we deal with this topic in Part 2).

With this in mind, we have approached the inquiry directly through the following question:

Will Facebook, YouTube, and Twitter be able to prevent or mitigate the next objectionable content crisis?

Key findings and assessment

OVERVIEW

During and after the March 15 terror attack, there was a rapid upload and transmission of a wide variety of digitally novel objectionable content within a very short space of time. We refer to this as an *objectionable content crisis* (abbreviated to "OCC" hereafter) to reflect that the quantity, variety, and frequency of uploads created an extreme outlier scenario.[3] Individual users acted separately and in coordination to exploit the time delay inherent in the content moderation process in order to continue uploading attack-related content, and ultimately move faster than content moderation processes could move.[4]

- At the time of the 15 March terror attack, all the platforms had content moderation processes in place for limiting the spread of objectionable content.
- These processes rely on a variety of mechanisms to detect, categorise (or classify), and database new objectionable content after it is uploaded or transmitted within the platform for the first time.
- After classification, automated systems can be used to automatically prevent attempts to upload or transmit identical copies of the information. However, there is a delay between the point in time that objectionable content is first uploaded to the platforms, and the point in time at which identical copies of that content can be automatically removed or prevented from being published.
- This delay is the main vulnerability in the platforms' ability to prevent the mass dissemination of objectionable content. This means the most important changes for preventing or mitigating the scale of future objectionable content crises will be changes that reduce the time delay between first upload and categorisation. Changes which do not reduce this time delay or enhance the accuracy and reliability of content classification are unlikely to be measurably effective.

KEY POINTS ABOUT CONTENT MODERATION AT SCALE

The investor group should understand the following about how content moderation is conducted at scale in digital platforms:

- When a digitally novel piece of content is uploaded for the first time, it is difficult (and sometimes impossible) to accurately detect and remove it using only automated tools. Detecting new objectionable content still requires human involvement most of the time.
- These humans (including both platform users and employees) cooperate to locate possibly objectionable content and assess it – taking account of its contents and context.
- After this assessment, the platform content moderators may then categorise or classify the content as objectionable in some way. If an objectionable classification is

[3] We do not rule out that such a scenario could have been anticipated, but we note the repeated comments made by the New Zealand Royal Commission of Inquiry that illustrate the attacker's commitment to "operational security" and his notable determination and commitment.

[4] We note the platforms took a range of unprecedented actions to respond, including by suspending ordinary content moderation processes (YouTube) and by using novel detection methods, such as detecting resemblances in audio rather than video (Facebook).

given to the content, the content (and the users transmitting it) may then be subject to one or more moderation actions. This may include creating a "digital fingerprint" of the content, known as a "hash", then uploading this fingerprint to a database.

- Only after the content is classified according to content moderation standards will the platforms' automated tools be able to prevent and remove all other digitally identical versions of the content from being uploaded or transmitted within the platforms. This whole process takes time to do accurately.

In the case of the 15 March terror attacks, copies of the original video were maliciously modified and distributed to create digitally novel content that would not be detected by the automated moderation systems.[5] In addition, users and organisations engaged in news reporting were also creating and sharing digitally novel versions of the video, even though they may have had no malicious intent. This made it difficult to constrain the spread of objectionable content during such a condensed period, particularly in combination with the increased levels of internet traffic and user attention focused on the 15 March terror attack.

To be clear, the platforms can efficiently locate, block, and remove content at scale once it has previously been classified as objectionable and logged in their databases.[6] However, content that is digitally novel and has never been classified cannot be detected and moderated at the same speed. When the Christchurch attacks occurred, the original livestream had not been classified, and neither had any of the hundreds of modified versions that appeared subsequently.

CHALLENGES OF MODERATION UNDER CRISIS CONDITIONS

We reiterate that the essential question for this assessment is whether the platforms have materially improved the speed and accuracy of their processes for detecting, categorising, and databasing novel objectionable content, particularly under crisis conditions. More specifically, the question is whether these improvements have sufficiently reduced the moderation time delay to withstand OCC conditions. Reducing this time delay is the critical outcome that will mitigate the scale of future content dissemination events similar to the 15 March 2019 OCC, even if it cannot prevent them entirely. In summary, we conclude:

- The measures introduced by the platforms have a high likelihood of significantly mitigating the scale and extent of future objectionable content crises.[7] They have taken steps to increase the speed and coordination of their shared responses, allowing them to more quickly detect new content and accurately classify it as objectionable, and then apply automated means to detect and remove it, as well as block identical copies at upload.

[5] The original livestream video was deliberately turned into memes by online communities. See the following for some examples: Wegener F, 'How the Far-Right Uses Memes in Online Warfare' (GNET) <https://gnet-research.org/2020/05/21/how-the-far-right-uses-memes-in-online-warfare/> accessed 13 April 2021.

[6] The public frequently questions why the platforms can act on copyright content so much faster than other kinds of content. This is because certain types of copyright content have already been classified. It is thought that platform systems for identifying copyright infringement are based on similar "hashing" techniques to those deployed through the GIFCT shared hash database.

[7] By "mitigate" and "scale", we mean reducing the total number of people who access or are exposed to objectionable content via the Platforms. Nevertheless, there remains a small risk that this number may be high in uncommonly bad cases.

- The platforms are highly unlikely to absolutely prevent the next objectionable content crisis.[8] It is difficult to classify content using automated tools and techniques, whether at the point of upload or at any point afterwards. Once new content has been uploaded there is an unavoidable delay before it can be accurately classified as objectionable, particularly if the platforms are expected to apply human rights principles that require the balancing of potentially competing interests. The platforms cannot eliminate this time gap entirely. Further, most current automated systems require a degree of human supervision (both as a matter of pragmatism and principle). The best the platforms can do during an OCC is to detect and classify novel content as quickly as possible after it appears on one of the platforms, then apply digital systems to automatically take down identical copies automatically thereafter.
- There is effectively no measure the platforms can introduce that could entirely prevent user exposure to violent objectionable content until that content has first been classified. While they play a critical role, automated systems offer imperfect solutions with significant trade-offs.[9] Even state of the art automation systems cannot currently identify which content is objectionable with total accuracy, and such advancements are not a realistic prospect in the foreseeable future. Automation can and should be improved, but all plausible automation systems that try and categorise new content will still have an error rate – e.g they are likely to remove some amount of non-objectionable content, while also erroneously permit some amount of objectionable content. The degree of error that is acceptable in either direction is a question of trade-offs between competing socio-political values.[10]
- The platforms continue to make reasonable efforts to reduce the extent of future objectionable content crises, given the measures they have taken in response to the complexity of conducting content moderation processes at scale under OCC conditions. The platforms' most effective changes focus on the essential problem identified above – the classification time delay. Some of the other changes made by the platforms are likely to exert only a weak or indirect influence on the core problem, and are otherwise unlikely to have any immediate measurable effect on preventing or mitigating a future OCC.
- It is generally sensible to assume that more funding and resources directed at any strategies that reduce the classification time delay will help to mitigate future OCC. This includes research into automated tools for detection and classification, increasing the total number of human moderators employed, and system-wide innovation – like partnerships with diverse community groups that can expedite and improve the accuracy of the content moderation process.[11] While the platforms have all increased their funding and resources for content moderation processes, we cannot say what amount of funding is reasonable, in part because this assessment rests on matters that cannot be known: for example, we cannot predict the likelihood of

[8] This takes "prevent" to mean that no person will be exposed to objectionable content produced during a real-world attack. We note from advisory materials provided to us by GSNZ that the institution's "long term goal" is "No further online sharing of objectionable content through the platform." This is a practically impossible standard to meet.

[9] The automation has a high error rate of both false positives and false negatives. This often has discriminatory effects.

[10] We explore this question of trade-offs in use of automation in more detail in Part 2, to the extent that regulation requires the use of automation in order to achieve compliance with content standards.

[11] We note that the importance of having diverse experience and expertise within content moderation teams has been emphasized by the UN Special Rapporteur on the Right to Freedom of Expression.

scientific or technological developments in automation research; we also cannot accurately compare the current state of content moderation as a whole with any hypothetical future state, in part because the current state is opaque and difficult to investigate. The same is true for the total number of content moderator employees. Questions of sufficiency and reasonableness will also depend upon personal perceptions of the wider problem area (specifically, the proliferation of "harmful content") and which values should be prioritised when managing trade-offs.

POPULAR MISCONCEPTIONS ABOUT CONTENT MODERATION AND LIVESTREAMING

The platforms might have been better prepared for an event like the 15 March terror attack. However, some misconceptions persist around how they may have contributed to the scale of the event. It is important to resolve these at the outset so that they are not given undue weight, since changes focused on these areas may have little effect on the extent of future OCCs:

- There is little to indicate that objectionable content linked to the terror attacks spread on the platforms because of inadequacies in the way content moderation guidelines were set. There is no doubt that the 15 March objectionable content breached the content standards of all three platforms.[12] Moreover, content did not spread on the platforms because of lenient attitudes toward freedom of speech or expression.
- An important feature of the OCC that followed the 15 March terror attack was the way online communities deliberately took steps to circumvent automated content moderation systems to enhance the content's spread. The platforms have taken steps to deal with such behaviour, however we anticipate that determined actors and the online communities that support them will continue to exploit the platforms' content moderation systems in whatever way they can discover. This is a known feature of online right wing extremist behaviour, and internet criminality generally. Moreover, the huge amount of legitimate news coverage of the event will also have resulted in the proliferation of new copies of the video. It is not clear how to prevent such news coverage legally or technically, or whether it is desirable to do so.

Similarly, many people assume that the original livestream of the 15 March terror attack was the major contributor to the OCC, and underestimate the role of the many subsequent videos and other files that were derived from the livestream. More specifically:

- The original livestream made a smaller contribution to the scale of the OCC in quantitative terms than is popularly believed. Facebook has shared data illustrating that the livestream was seen by far fewer people than subsequent content produced from the livestream. Most dissemination of objectionable content was achieved through the subsequent upload and download of non-livestreamed video files. Despite this, much policy, regulatory, and legislative attention has been unduly directed at the activity of "livestreaming" (including explicitly within the Christchurch Call).
- Similarly broad dissemination of objectionable content probably could have been achieved without livestreaming, through conventional upload of a video file to a cloud file hosting site.[13] It is even plausible that the perpetrator could have recorded

[12] This point is made forcefully in: Douek, E 'Australia's "Abhorrent Violent Material" Law: Shouting "Nerd Harder" and Drowning Out Speech' (Social Science Research Network 2019) SSRN Scholarly Paper ID 3443220 <https://papers.ssrn.com/abstract=3443220> accessed 6 April 2021.

[13] Links to this could then have been distributed online through the use of scripts, accomplices, or both.

the video to his mobile device, paused during the attack, and uploaded the video to a file sharing website for further distribution.

- Nation states, media, and the public frequently conceptualise the 15 March objectionable content as "the video", "the livestream", or words to that effect. This creates the impression that there was only a single digital artefact for the platforms to classify and moderate. In fact, there were hundreds of unique videos, each containing thousands of objectionable frames (individual images that could be distributed separately).[14]
- The 15 March OCC featured the rapid creation, upload, and dissemination of large volumes of non-livestreamed video files. This attacked the same fundamental weakness as the livestream itself – the relatively brief but unavoidable time delay inherent in the process of detecting, classifying, and databasing new content.

Which measures will be most effective at mitigating an OCC?

The most effective means of dealing with violent extremist material during an OCC come from cross-platform collaboration efforts. Some of these efforts are based around collaboration with other multilateral international bodies, including the United Nations and the European Internet Forum. Most of these multilateral bodies pre-date the Christchurch attacks. Twitter, Facebook, and Google/YouTube are core partners, founders, funders, members, and supporters of many of these collaboration efforts.[15]

The following mechanisms and interventions exact the most direct mitigating (and to a lesser extent preventative) effect on the scale of any future OCC. These interventions merit the most attention when advocating for further improvements and transparency.

DEVELOPMENT OF GIFCT CONTENT INCIDENT PROTOCOL ("CIP")

In accordance with the Christchurch Call, the Global Internet Forum to Counter Terrorism (GIFCT) has introduced a new Content Incident Protocol. The CIP was developed as a response to the events of the 15 March terror attack and resulted from a commitment made by the platforms under the Christchurch Call. The protocol aims to thwart the online proliferation of content produced by a perpetrator during a real-world attack.

When a CIP is declared, the platforms coordinate to rapidly classify content produced by a perpetrator or accomplice. Once content has been classified, "hashes" — unique digital "fingerprints" — are rapidly added to a shared database. As part of a CIP, continuous communication is also established between the platforms:[16]

> *By declaring a CIP, all hashes of an attacker's video and other related content is shared in the GIFCT hash database with other GIFCT member platforms. Furthermore, a continuous stream of communication is established among all GIFCT founding members to identify and address risks and needs during an active CIP. The CIP is a multi-step process, including a decision to activate the CIP, communication of that decision, a review of content assets,*

[14] 6.8% of hashes in the GIFCT shared hash database relate to the Christchurch attacks. Hashes from two other attacks where a CIP was activated represent 2% and 0.1% respectively.

[15] Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism. (n.d.). Blog.Youtube. Retrieved 1 April 2021, from <https://blog.youtube/news-and-events/facebook-microsoft-twitter-and-youtube/>.

[16] Crisis Response. GIFCT. Retrieved 5 April 2021, from <https://gifct.org/crisis-communications/>.

and other steps, to inform GIFCT member companies and relevant governments about content from the real-world event that may be manifesting online. A CIP ends with an official "conclusion" determined by impacted GIFCT platforms once the volume of content has noticeably decreased.

In what situations is a CIP activated?

The circumstances in which a CIP is activated are tightly constrained. The precise contents and boundaries of the protocols are strictly confidential, to prevent them from being exploited. This confidentiality makes external analysis of those protocols difficult and creates a tension between preventing exploitation and the value that greater transparency around the CIP would provide. We note that the parameters of the CIP have been tested in at least two sets of tabletop exercises with government and non-government participants.[17]

The CIP is only activated after an initial assessment process has been followed. This assessment process has been initiated 100 separate times between March 2019 and November 2020.[18] Of these, the CIP itself has been activated twice: first in response to the 9 October 2019 terrorist attack in Halle, Germany, and later in response to the 20 May 2020 terrorist attack in Glendale, Arizona.[19] The activation of the CIP in these two cases should reassure the investor group that the CIP can and does play a role in limiting the impact of some OCC. We are persuaded that the CIP played a role in suppressing the objectionable content of these two events, although we note that the volume of content produced by both news media and adversarial online communities appears to be drastically smaller than occurred around 15 March.[20]

Limitations of the CIP

The CIP may not be activated for all content that the investor group considers objectionable, including situations where a perpetrator livestreams audio-visual content during real world violence. An example of this is the Nakhon Ratchasima shootings of 8 and 9 February 2020, during which a soldier of the Royal Thai Army killed 30 people and wounded 57 others. During these attacks, the perpetrator livestreamed to Facebook intermittently. In the absence of further information, the non-activation of the CIP suggests that the criteria for CIP activation are narrower than "content produced by a perpetrator during a real-world attack". It is likely that to meet the CIP criteria, the content produced by a perpetrator must itself depict actual on-screen acts of physical violence, where (to our knowledge) the Nakhon Ratchasima livestream did not.[21] It also appears the CIP assessment process involves an assessment of how likely the content is to be virally spread.[22]

[17] Tom Barraclough from Brainbox participated in exercises testing these protocols in Wellington, New Zealand, in December 2019.

[18] Crisis Response. (n.d.). GIFCT. Retrieved 5 April 2021, from <https://gifct.org/crisis-communications/>. We were unable to identify any more up-to-date figures.

[19] 'GIFCT Transparency Report, July 2020' <https://gifct.org/wp-content/uploads/2020/10/GIFCT-Transparency-Report-July-2020-Final.pdf> accessed 5 April 2021.

[20] GIFCT has shared percentages to indicate what proportion of content in the total shared hash database relates to each attack: Christchurch perpetrator content represents 6.8%; Halle, Germany perpetrator content represents 2%; Glendale, Arizona Perpetrator Content represents 0.1%.

[21] We note that Facebook suspended the perpetrator's account, thus removing objectionable content by means a content moderation different mechanism.

[22] 'Christchurch Call Community Consultation: Final Report' (2021). <https://www.christchurchcall.com/christchurch-call-community-consultation-report.pdf> accessed 14 April 2021.

In addition, a CIP is unlikely to be activated for footage of real world violence that has been captured by bystanders to an attack. As far as we know, the CIP will only be activated in response to content created by the perpetrator or an accomplice. Recording by bystanders may depict the same violent acts as recordings by a perpetrator,[23] and may have similar traumatic effects on some viewers. Bystander recordings may also amplify the terrorist's goal of publicising an attack, regardless of the bystander's intent.

Questions around the distinction between perpetrator-led and bystander-led footage frequently arise for the platforms.[24] It is not always clear whether bystander footage is, should, or will be classified as objectionable. Our best assessment is that each instance of bystander footage is classified on a case-by-case basis. Even where emergency protocols are not activated, content is still being moderated according to usual processes. If bystander content is flagged by manual or automated systems and assessed as infringing the platforms' policies, it will be dealt with accordingly. We raise this aspect of the overall subject because the restriction of emergency response protocols to perpetrator-or-accomplice footage is a potentially significant limitation on whether those protocols limit the circulation of content that might otherwise be objectionable.[25]

GIFCT SHARED HASH DATABASE

GIFCT has a shared database of "hashes", or digital fingerprints of harmful content which has previously been classified by the platforms as being "objectionable" in the manner articulated in our brief. The database is now accessible by 13 different technology companies (expanded to 14 in July/August 2021).

The shared hash database was initiated before 15 March 2019, primarily in response to religious extremist terrorist attacks in Europe across 2015 and 2016. GIFCT was initially founded in June 2017 to house the shared hash database before becoming an independent organisation with an Executive Leadership team and a 24-hour crisis response team — a move that was accelerated through the Christchurch Call.[26]

The shared hash database is a system for preventing the upload of content that matches the digital fingerprint of content previously classified as objectionable. Once classified, it can also be used to remove objectionable content which has been previously uploaded that matches the hash. The database is made available to members of GIFCT to incorporate into their content moderation systems in whatever way they wish, and new platforms can gain access to the database by joining GIFCT. The shared hash database is highly effective at identifying digital duplicates of infringing content.

The database has two broad limitations, both of which present trade-offs:

1) Institutional limitation: we identified repeated expressions of concern by commentators about the oversight, transparency and auditability of this database, which is

[23] Two recent examples are: the CCTV footage from inside the Christchurch Mosques; and bystander footage of the aftermath of the shootings in Boulder, Colorado.

[24] Australia's Abhorrent Violent Material draws a similar distinction, as does the CIP.

[25] For example, consider the fact that there is a CCTV recording depicting the 15 March attacks from inside the Al Noor mosque. It is not clear that the CIP would cover this in the event that this recording was uploaded to the Platforms.

[26] 'Christchurch Call Community Consultation: Final Report' (2021) <https://www.christchurchcall.com/christchurch-call-community-consultation-report.pdf> accessed 14 April 2021.

controlled by GIFCT. For example, there is no clear mechanism for challenging the decision to add a piece of content to the database. Other concerns related to consistency and transparency around the criteria for adding content to the database. Because the database can be so effective at removing infringing content, this can lead to harm if content is wrongfully removed rapidly, without notice and at scale. Initially, the database was not to be used for automated takedowns, but it appears this is increasingly how it is used by the companies.

2) Technological limitation: it is not clear how far the shared hash database can deal with modifications to content that has been previously hashed. This is a significant limitation of conventional hashing techniques. The database has substantially less utility if it only identifies exact duplicates. For example, the 800 digitally unique variants of the 15 March livestream video would likely not have been identified by the hash database unless perceptual hashing techniques were used. On the other hand, if the database was limited only to exact duplicates, this would lower the risk that content is wrongfully taken down because of broad but ultimately incidental visual similarity to banned content.

At its annual summit in July/August 2021, GIFCT announced a range of new initiatives to improve its shared hash database. GIFCT also released a widely acclaimed independent human rights impact assessment of its organisation which it had commissioned, although a detailed analysis of this assessment is beyond the scope of this brief.[27]

ONGOING RESEARCH INTO PERCEPTUAL HASHING TECHNIQUES

Hashing is an area of ongoing technological research and development. The platforms are supporting some of this research.[28] As part of this work, some new hashing techniques that use image data may help capture content in situations where modifications have been made, but the content is visually similar to content that has been previously hashed. These techniques are referred to as "perceptual hashing". We located research suggesting that the GIFCT hashing database may already make use of perceptual hashing techniques, but cannot confirm this given the opacity of the database.[29]

There may be trade-offs to adopting perceptual hashing techniques in the GIFCT database. In particular, any digital system which assesses visual similarity between two pieces of content may lead to a higher rate of false positives or false negatives. This risk is not present if the hashing database is only comparing content at the digital level to identify exact duplicates.

[27] Available from: <https://gifct.org/2021/07/20/a-human-rights-based-approach-to-preventing-terrorist-and-violent-extremist-exploitation-of-the-internet/>.

[28] Facebook for example has "open sourced" some of its hashing techniques.

[29] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945. <https://doi.org/10.1177/2053951719897945>.

The comparative strengths and limitations of the GIFCT shared hash database are a useful illustration that:

- a range of techniques will be required when it comes to automated content moderation systems;
- there will be trade-offs in how these systems operate; and
- appropriate human oversight and auditing of those socio-technical systems will be important for ensuring ongoing trust and confidence in those systems.

DEVELOPMENT OF CHRISTCHURCH CALL AND EU CRISIS RESPONSE PROTOCOLS

After the 15 March terror attacks, GIFCT and various governments collaborated to formulate two shared crisis response protocols. The first protocol relates to the Christchurch Call and the second relates to the European Union. These response protocols are complementary to the Content Incident Protocol (CIP), but remain separate. The CIP is restricted to the tech platforms themselves.

The full details of the Crisis Response Protocols are confidential. In general, they focus on formalising standards of what constitutes an “incident”, implementing a response process for such incidents, and opening communication lines between the platforms and with government actors in order to make any response more effective. There is little publicly available data about the Crisis Response Protocols, including when they have been followed or not.

As a supplement to the narrower CIP, the Crisis Response Protocols will play an important part in responding to any OCC that does not meet the threshold for CIP activation.

We also note that the platforms continue to run their normal content moderation processes, irrespective of whether measures like the CIP or Crisis Response Protocol are engaged. For example, Facebook independently removed the account of the Nakhon Ratchasima perpetrator, including the livestream and all other content he had generated.

Greater transparency around the Crisis Response Protocols would be useful, however we acknowledge that it is difficult to facilitate access and auditing of these protocols by external parties because of their role in responding to crisis conditions and their role in law enforcement.

The Christchurch Call Shared Crisis Response Protocol is opted-into by signing on to the Christchurch Call. To our knowledge, the protocol is not applied to countries who are not signatories. This decreases the protocol’s effectiveness to the extent that it may not act as a measure to suppress the distribution of objectionable content in some countries. We are not in a position to suggest that membership of the Christchurch Call or access to its Crisis Response Protocols should be expanded. That is because access to the Call and the protocols is limited to countries that sign up to the values of the Christchurch Call. If countries are unable to demonstrate a commitment to these values, then the Call’s values probably should not be compromised solely in order to expand membership.

Less direct measures for mitigating an OCC

In our assessment, the multi-stakeholder collaborative efforts outlined in the section above are most likely to be effective at preventing or mitigating an OCC. In this section we describe a range of other measures that could be relevant to investor decision-making and do play some role in mitigating an OCC. However, our conclusion is that these measures play a less direct role in responding to the unique conditions faced during an OCC.

CONTENT MODERATION SYSTEMS: ONGOING REFINEMENTS, STUDY, AND RESOURCING

Content moderation by the platforms is necessary but not sufficient for responding to an OCC. At the time of the attacks, each of the platforms deployed a variety of content moderation systems to limit the appearance of objectionable content on its application or website. These systems involved both humans and computer systems working in cooperation to moderate content.

Before content can be acted upon during an OCC, it must first (1) come to the attention of content moderators through manual or automated "flagging" processes, then (2) be classified according to relevant policies. Only then (3) can any duplicate or sufficiently similar material be automatedly identified, matched, and subject to a moderation action.

Many of the initiatives announced by the platforms since the OCC contribute to the enhancement of those three steps in the content moderation process. These include improvements to the resourcing, technologies, or procedures that relate to:

- human processes of content moderation;
- technological processes of content moderation;
- clarifying policies applied during content moderation or revising those policies to alter their intended effect; and
- implementing triage mechanisms which bring content to the attention of moderators more quickly.

While broad improvements to content moderation systems will play a fundamental role in preventing or mitigating an OCC, they are insufficient during such extreme circumstances. Because of these findings, we have concluded there is limited merit to cataloguing and analysing each change to content moderation systems made by the platforms that may have some bearing on responding to an OCC, and have not included such a catalogue in this report. There are also a range of pragmatic reasons, listed below, why such an assessment would be of limited value. We explain these reasons because they also illustrate why the regulatory approaches we recommend in Part 2 are desirable, because they would facilitate this kind of cataloguing exercise:

- To catalogue such changes effectively would require accounting for minor changes to content moderation policies and practices which are not publicly available to outside groups for scrutiny.
- To catalogue each change would require us to account for the wide range of improvements being incorporated through the platforms' research in machine learning and other automated techniques. In reality, each platform uses multiple different machine learning models to recommend and moderate content.

- There is no way of forecasting the impact of many such changes, which frequently are not solely related to an OCC, as the core subject of our investigation.
- External transparency reporting does not assist and the volume of such reporting is significant. The platforms frequently offer metrics to indicate how much content they are removing, and what proportion of those removals occur before any person has been exposed to that content. Unfortunately, this information is often unhelpful without further context.
- Even if sufficient data could be identified to understand how content is being moderated, investigators would need a previous baseline against which that data could be assessed. It is difficult to formulate a baseline for assessment when the scope of the platforms' policies is constantly changing.
- Equally, it is difficult to assess the platforms content moderation efforts by comparing them with other bodies (for example, other platforms, or other content moderation and censorship bodies). The scale of the companies is enormous and their products are unique (noting many of the platforms have more than one product requiring content moderation). It would be difficult to find reasonable comparison organisations whose performance could act as a comparative standard for assessment.

Accordingly, cataloguing such changes cannot be reasonably completed within the scope of the existing project, and regardless, would be of limited value. Nonetheless, we acknowledge that the platforms have made important efforts to increase the transparency of their content moderation interventions.

If the investor group wishes to monitor the effectiveness of content moderation changes over time, the best ways of doing this are by seeking out existing audits of platforms' self-assessments through independent bodies, or critical analysis by independent researchers.[30] This is one reason why, in Part 2, we endorse regulatory approaches which enhance transparency and standardise reporting on how content moderation is being conducted.

It is important to bear in mind that much of the public commentary about failures in content moderation relate to individual cases: these are only a loose indicator of whether the platforms are moderating content correctly according to their policies at scale. Further, they are often reported in news media or on social media in ways that do not give a full appreciation of either the relevant policies involved, or the full facts of the individual case, nor whether complaints and reporting processes were followed by the complainant. We therefore caution against pointing to individual failures of content moderation systems as being evidence of wider systemic flaws without careful investigation.

THE IMPACT OF FURTHER RESEARCH AND RESEARCH FUNDING

All the platforms have committed significant research funding toward technical and social methods of countering violent extremism, and towards countering extremist use of technology platforms. In addition, the platforms are heavily invested in the development of artificial intelligence techniques that might enhance their ability to automatically detect and classify content with greater accuracy and speed.

[30] We note the role of the assessment processes through the Global Network Initiative and the EU Code of Conduct on Illegal Hate Speech, discussed elsewhere in this report.

GIFCT is also undertaking research through research partnerships with its own research network (GNET), and other independent research bodies. While we are unable to offer meaningful comment on whether the overall level of funding contributed is adequate, we note that this funding has led to a substantial body of research.

We observed frequent calls to improve researcher access to the platforms for research purposes. The investor group could consider adding its support to these calls for enhanced research access.

COMMENT ON POLICIES THAT LIMIT ACCESS TO LIVESTREAMING SERVICES

Both Facebook and YouTube implemented new policies that restrict access to livestreaming on their platforms under certain conditions. We have low confidence that these measures will prevent future OCC, materially reduce the scale of a future OCC, or prevent the platforms from otherwise being exploited to disseminate objectionable content during an OCC. In particular, we draw attention to comments made by Facebook's Head of Global Policy, Nick Clegg, regarding Facebook's "one strike" livestream restrictions. At the 2019 Christchurch Call summit meeting in Paris, Clegg said that**:**[31]

> *Those restrictions, if they had been in place at the time of the Christchurch atrocity, would have prevented the terrorist from using his live account on that day.*

We found no information in the report of the New Zealand Royal Commission of Inquiry into the attacks to suggest that any of the individual's activity on Facebook was found to have breached Facebook policies in the lead-in to the 15 March terror attacks.[32] We cannot see how Clegg's claim can be correct unless it is based on information that is not publicly available – for example, information held by Facebook that the individual was sanctioned for a breach of serious policy in the weeks or months preceding the attack. Unless such a sanction occurred, the one strike policy would not have prevented the individual from making use of Facebook Live.[33]

As such, it is not possible to conclude that the new Facebook Live policy would have prevented the terrorist's ability to livestream the attacks as he did. Subsequently, we conclude that the policy would not prevent a future OCC in materially similar circumstances, though it may have a preventative or mitigating effect in different circumstances.

[31] 'Facebook Says New Rule Would Have Stopped Christchurch Shooter Livestreaming' (Stuff, 15 May 2019) <https://www.stuff.co.nz/national/politics/112756865/facebook-says-new-rule-would-have-stopped-christchurch-shooter-livestreaming> accessed 14 April 2021.

[32] We note the Royal Commission also closely analysed the individual's Facebook activity and concluded that his activity would not have justified escalation by either New Zealand intelligence agencies or Police.

[33] This new policy effectively means that a breach of any of Facebook's "most serious policies" results in the user losing access to the Facebook Live service for a specified period: "We will now apply a 'one strike' policy to Live in connection with a broader range of offenses. From now on, anyone who violates our most serious policies will be restricted from using Live for set periods of time – for example 30 days – starting on their first offense. For instance, someone who shares a link to a statement from a terrorist group with no context will now be immediately blocked from using Live for a set period of time."

The preventative effect of any policies that restrict access to livestreaming functions will be diminished by the following:

- It is possible that a potential perpetrator would be forewarned (by communication from Facebook informing them of their policy breach) that they are temporarily barred from livestreaming, and for what period of time.
- A sufficiently determined perpetrator can defer their attack until after a temporary ban has expired.
- Many perpetrators will probably avoid committing any action that will breach a serious policy, and therefore will never be banned from the use of Facebook Live. For comparison, the Royal Commission noted the 15 March terrorist's commitment to operational discipline, including conducting online activities in a way that would avoid suspicion or detection.

According to the Royal Commission, the 15 March terrorist had other popular livestream-capable applications installed on his mobile device.[34] Using these, he could have livestreamed elsewhere, with non-livestream versions of the video then able to be uploaded to the platforms for further dissemination in exactly the same way as occurred. We understand that the attacker used an intermediary application (intended for capturing user-generated sports videos) to record the video, rather than Facebook's native livestreaming capabilities. In short, while livestream restrictions may prevent or delay some future attackers from livestreaming on the platforms, they do not prevent the upload of non-livestreamed copies of objectionable videos, or the use of the platforms to host links to livestreams on other parts of the internet.

When it comes to helping investors assess their position on restrictions on access to livestreaming, it is important to consider the human rights implications arising from any situation where livestreaming is pre-emptively restricted without first demonstrating a violation justifying that restriction. We do not endorse policies that restrict access to livestreaming arbitrarily. We expand on these factors in Part 2.

Trade-offs to consider when calling for further action

We conclude that the platforms are making reasonable efforts to mitigate the success of future OCC. However, we are aware that the investor group may wish to call for further action. In this section, we describe broadly what those further actions might be and briefly outline the trade-offs introduced by advocating for such actions.

CALLS FOR INCREASING AUTOMATED CLASSIFICATION IN CONTENT MODERATION

The platforms use automation in several different ways for content moderation purposes. In the simplest sense they use it for matching digitally identical content with content that has previously been classified as objectionable by a human moderator, so that all identical copies can be blocked at upload thereafter by the automated tools.

[34] Report of the Royal Commission of Inquiry into the Terrorist Attack on Christchurch Mosques on 15 March 2019 at p 228.

This kind of automation has a relatively narrow function for which it is highly effective. A major limitation of such tools is that they cannot act upon a novel piece of content until it has appeared on one of the platforms and been classified by a human moderator. In this sense, they are inherently reactive tools. Somebody, somewhere must see the new objectionable content after it is uploaded.

However, the platforms also deploy automated tools for more complicated moderation tasks. In particular, the platforms use automated tools that *classify* new content as objectionable. Theoretically, these kinds of tools can prevent novel objectionable content from ever being uploaded to the platforms. Alternatively, even where such tools may not prevent initial upload, they still can expedite the classification time-delay by filtering for new content which probabilistically might be objectionable, so that it can receive faster assessment by a human content moderator.

This latter kind of automation – which we loosely refer to as tools of classification – offers greater preventative potential, but at the price of significant trade-offs. Chief among these trade-offs is the risk of inaccuracies. Inaccuracy may occur in the form of false negatives – i.e., the tool fails to detect that a novel piece of content is objectionable, and thus allows it onto the platform. This reduces both the reliability and the preventative potential of the automation, though it still may be useful in conjunction with other content moderation methods that utilise human judgment.

Inaccuracy may also occur in the form of false positives – i.e., the tool erroneously treats non-objectionable content as if it is objectionable, resulting in it being blocked at the point of upload or automatically removed at some later point. This trade-off is arguably more pernicious. It has significant human rights ramifications, and can undermine aspects of the platforms’ core value to their users.

The simplest explanation for why these trade-offs are difficult to mitigate is that it is difficult for computer systems to accurately classify content, especially where content moderation standards require human judgement, or are vague or ambiguous. Accurate classification generally requires a nuanced judgment of the semantic content of audio-visual documents, including their meaning in context. Automated tools are currently poor at this, and it is even harder for them to appropriately weigh the context in which content is being shared, commented on, critiqued, etc.

In short, while improved automation is a justifiably important part of the future of content moderation systems, there are serious risks to asking the platforms to do too much with tools that are inadequate for the purpose. Furthermore, automated systems for content moderation that use artificial intelligence techniques are subject to the same kinds of considerations as any AI system. In particular, they may be biased in ways that reflect the social or other biases of the people creating them, and may have discriminatory effects.

The key takeaway for investors is that automation and the use of algorithmic systems are inevitable and necessary for the platforms to operate. However, automation has flaws: it requires appropriate human oversight and should not be uncritically endorsed given the risks and trade-offs it creates. To the extent that regulatory approaches require the use of automated content moderation techniques in order to remove content rapidly, this creates significant risks of false-positive take-downs, or take-downs which are too rapid to account for complex contextual and legal considerations. There is wide consensus that some regulators are placing too much confidence in the platforms’ ability to use automated tools.

CALLS FOR GREATER TRANSPARENCY AROUND CONTENT MODERATION

At the outset, we note that some of the platforms have begun providing better transparency and access to some of their internal operations and statistics around matters material to prevention of OCC. Nonetheless, one of the most common criticisms of many of the measures taken by the platforms (including the GIFCT) is a lack of transparency. There are many instances where information, data, and statistics provided by the platforms may sound impressive at face value, but are meaningless without further contextualisation within wider datasets that are not made available for public scrutiny. This problem is summarised below by Evelyn Douek:[35]

> *Companies similarly appealed to the legitimacy of the GIFCT in the wake of the Christchurch massacre as evidence of their commitment to fighting the spread of violent footage. But when GIFCT members boasted that they had added over 800 new hashes to the database, there was no way to verify what this meant or whether it was a good marker of success. There was, for example, no way to know if these included legitimate media reports that used snippets of the footage, or completely erroneous content, or what proportion of the variants of footage uploaded the figure represented. These deficiencies repeated themselves in the wake of the Halle livestream, even as the platforms were congratulated for their effective response.*

This does not mean that the platforms are totally opaque. They allow themselves to be audited by external institutions: specifically, the Global Network Initiative, and the EU Internet Forum. The GNI is a network of companies that commit to a set of principles, against which they are assessed every two years.[36] These principles are comprehensive, but by way of illustration, they include the following:

- “All human rights are indivisible, interdependent, and interrelated: the improvement of one right facilitates advancement of the others; the deprivation of one right adversely affects others. Freedom of expression and privacy are an explicit part of this international framework of human rights and are enabling rights that facilitate the meaningful realization of other human rights.”
- “The duty of governments to respect, protect, promote and fulfil human rights is the foundation of this human rights framework. That duty includes ensuring that national laws, regulations and policies are consistent with international human rights laws and standards on freedom of expression and privacy.”
- “ICT companies have the responsibility to respect and promote the freedom of expression and privacy rights of their users. ... The collaboration between the ICT industry, investors, civil society organizations, academics and other stakeholders can strengthen efforts to work with governments to advance freedom of expression and privacy globally.”

The GNI principles also refer to and rely on the importance of: the right to privacy; the right to freedom of expression; the importance of responsible decision-making by companies; and the importance of multi-stakeholder collaboration. The principles also emphasise the importance of governance structures that support the purpose and implementation of

[35] Douek, E ‘The Rise of Content Cartels’ [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3572309> accessed 31 March 2021.

[36] The GNI has been cited with approval by the UN Special Rapporteur for Freedom of Expression with companies advised to engage with the GN’s processes.

the principles, hold company governance accountable, and demonstrate compliance with the principles through independent assessment and evaluation, and systems of transparency with the public.

The GNI assessment process itself is confidential, but there are a range of reports available documenting those assessments. We note that the Electronic Frontier Foundation has rejected these audits as insufficient, given that they failed to disclose the platform-government collaborations which were later revealed by Edward Snowden.[37]

The EU also has a Code of Conduct on Countering Illegal Hate Speech Online, which relates to the establishment of a body called the EU Internet Forum in December 2015. The Code was agreed upon Facebook, Microsoft, Twitter and YouTube in May 2016. The Code's implementation "is evaluated through a regular monitoring exercise set up in collaboration with a network of organisations located in the different EU countries. Using a commonly agreed methodology, these organisations test how the IT companies are implementing the commitments in the Code."[38] The fifth monitoring round concluded in June 2020. According to the EU, "the Code of Conduct is delivering continuous progress: the last evaluation shows that on average the companies are now assessing 90% of flagged content within 24 hours and 71% of the content deemed illegal hate speech is removed."[39] We note that the secrecy surrounding the development of this code has been criticised.[40]

In general, we support measures that enhance transparency and access to the platforms on matters material to the prevention and mitigation of future OCC. However, these must be balanced against the risk that individuals could use such information to circumvent the platform's content moderation systems. Perpetrators, accomplices, and a wider set of internet users can and will leverage transparency around the platforms' systems in ways that exploit the weaknesses in those systems. In particular, it is well known that extreme right-wing movements deliberately communicate online in ways that disguise, misdirect, or otherwise insulate them from detection and moderation.[41] Some of this communication is strategically designed to attack content moderation efforts and undermine them in the eyes of the public.

CALLS TO CEASE SERVICE IN A CRISIS

The scale of an OCC would be drastically mitigated if the platforms universally ceased providing services for the duration of an attack. There is a growing international trend where states block access to social media or to the internet during violent crises. For example, the Sri Lankan government blocked access to social media during a series of bombing

[37] 'GNI Resignation Letter'. Electronic Frontier Foundation, 9 Oct. 2013, <https://www.eff.org/document/gni-resignation-letter>.

[38] 'The EU Code of Conduct on Countering Illegal Hate Speech Online' (European Commission - European Commission) <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 1 April 2021.

[39] Ibid.

[40] The same author attributes the establishment of the GIFCT to the platforms' attempts to comply with the European code. See: Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945. <https://doi.org/10.1177/2053951719897945>.

[41] The Royal Commission's report notes how right-wing extremists use humour and irony to disguise their intentions. The shooter adopted such practices even in discussions with the Commission.

attacks.[42] However, this is not a proportionate response for many reasons and is opposed by human rights monitoring bodies.

Total cessation of service has a suite of significant trade-offs, most obviously that people use the platforms to communicate in times of crisis, and even to coordinate security and law enforcement responses. Cessation has not been seriously proposed by the platforms as a legitimate response to an OCC. Partial cessation of normal practice and procedure did take place in response to the Christchurch attacks. In particular, YouTube began to remove large quantities of content using automated systems without first assessing whether all of it could be justifiably classified as objectionable.[43] Another factor to consider here is the way that cessation of service in a crisis could undermine crisis response and prevent the creation of evidence necessary for investigation criminal activity. For these reasons, total cessation of services in response to an OCC raises significant trade-offs which are unlikely to be justified or proportionate.

CALLS FOR GREATER ADOPTION OF HUMAN RIGHTS PRINCIPLES IN CONTENT MODERATION

We frequently encountered the suggestion that the platforms should place a greater emphasis on human rights throughout their governance structure, policies, and operations, and that this would contribute to preventing or minimising the scale of an OCC. We agree that the platforms should adopt human rights approaches to content moderation. Importantly, however, "taking a human rights approach" must not be misconstrued as a synonym for total elimination of all objectionable content (however defined), or as a synonym for perfect content moderation according to a set of subjectively acceptable standards. We explain what it means to take a human rights approach in relation to content moderation in Part 2.

Human rights approaches do not just protect the victims of attacks. They protect all people, including perpetrators and individuals who share objectionable content online. The right to freedom of expression is a human right, as is the right to privacy and freedom of association. Human rights can also frequently conflict with each other, meaning reasonable people can reach different conclusions on how they apply.[44] The Facebook Oversight Board considers human rights instruments in its decisions and this has led at times to conflicting conclusions between Oversight Board members. Equally, all the platforms have some form of human rights input into their content moderation policies, processes or governance arrangements and are subject to the United Nations' Office of the High

[42] Amarasingam DA, 'Turning the Tap Off: The Impacts of Social Media Shutdown After Sri Lanka's Easter Attacks' (GNET) <https://gnet-research.org/2021/03/05/turning-the-tap-off-the-impacts-of-social-media-shutdown-after-sri-lankas-easter-attacks/> accessed 13 April 2021.

[43] 'The Christchurch Attacks: Livestream Terror in the Viral Video Age' (Combating Terrorism Center at West Point, 18 July 2019) <https://ctc.usma.edu/christchurch-attacks-livestream-terror-viral-video-age/> accessed 14 April 2021.

[44] On 14 April 2021, the Oversight Board released another decision dealing with the intersection between Facebook's content moderation guidelines and international human rights law related to "zwarte Piet" (Case decision 2021-002-FB-UA): "Numerous human rights are implicated in this case beyond expression, including cultural rights, equality and non-discrimination, mental health, and the rights of children. The Board seeks to evaluate whether this content should be restored to Facebook through three lenses: Facebook's Community Standards; the company's values; and its human rights responsibilities. The complexity of these issues allows reasonable people to reach different conclusions, and the Board was divided on this case."

Commission of Human Rights' Guiding Principles on Business and Human Rights.[45]

The key point is that a commitment to human rights does not always make content moderation judgments simpler or less ambiguous, and can make them more challenging to apply and slower in a crisis. This is one complication of regulatory proposals that we review in Part 2, particularly Germany's NetzDG and the Australian Abhorrent Violent Material amendments.

We also note that aggressive content moderation to remove objectionable content can have negative human rights impacts. Organisations focused on documenting human rights abuses, like The Syrian Archive and WITNESS, have conducted extensive international advocacy on the way that the automated removal of content can result in the destruction of evidence of war crimes and other state-led abuses. Such evidence might otherwise be used to protect human rights during prosecutions. Similar concerns can arise around the policies that curtail access to livestreaming services. For example, YouTube's new policy preventing any user with less than 1000 subscribers from livestreaming from a mobile device effectively prevents ordinary individuals from broadcasting contemporaneous video record of a human rights abuse in progress – including video like that which recorded the death of people like Philando Castile and George Floyd at the hands of police officers.

Human rights as a concept and legal device flow from their history as a tool to protect individuals from abuses by nation states.[46] This history is essential context for any suggestion that the platforms should be assisting nation states (whether voluntarily or by regulatory compulsion) to limit the human rights of individuals, for example by limiting freedom to express and receive information, limiting rights to privacy, or limiting rights of freedom of association. We touch upon this in greater detail in Part 2.

CALLS TO ACCELERATE REGULATORY INTERVENTION

There is a growing recognition that asking large private tech platforms to be global content moderators raises different kinds of trade-offs for freedom of expression and for the platforms' relationships with nation states, particularly given the platforms' relationship with violent extremist content, polarising political content, allegations of election interference, and mis- and disinformation. The platforms are increasingly calling for guidance from nation states in the form of regulation, and states are also increasingly discussing proposals to regulate.

In the long term, the platforms are poorly placed and ill-equipped to be making momentous decisions about what expression is acceptable or not. The platforms have adopted this position themselves. This, in part, is what has led to the development of the Oversight Board by Facebook, and Twitter's attempt to develop the de-centralised "Blue Sky" content moderation protocol. The question of how to moderate content at scale is very much an evolving matter of academic, political, and legal expertise. The development of this expertise, including required funding and access to the companies' platforms, is a matter that investors could support.

[45] A comprehensive assessment of the way that the platforms have incorporated human rights instruments into their processes is beyond the scope of this assessment.

[46] This is not to dispute that the companies are subject to human rights instruments and subject to obligations to protect human rights too, even though they are not nation states.

We caution against perceiving regulation as a panacea in this area. Even if regulators decide to set content moderation standards themselves, this will not simplify the core tasks required in content moderation, including detection, classification and intervention in content. If anything it may complicate this task by introducing a range of competing legal frameworks or vague and imprecise standards that generate legal and factual complexity.

In Part 2, we explain in greater detail why we believe that regulation mandating particular standards for what content must be moderated and how is unlikely to be of significant benefit. In particular, we conclude that, in order for platforms to demonstrate a genuine commitment to a human rights approach, they may be required to push back against states that attempt to regulate their behaviour and introduce regulatory instruments that compel them to act in ways that unjustifiably limit their users' human rights.

While platforms are not democratically well placed to declare standards of free speech, they are well placed to rapidly triage content of the kind that was disseminated around 15 March 2019, and they have implemented mechanisms to rapidly intervene in such cases: in fact, they can intervene much faster than any government body could. For this reason, they will always have a significant role to play in global content moderation and the mitigation of objectionable content crises.

CONCLUSION TO PART 1

Summary

Broadly speaking, the platforms have made reasonable efforts to mitigate the scale of future OCCs. These efforts are likely to be highly effective at mitigating the scale of an OCC, even though we do not think that all future OCCs can be prevented entirely. While effective, all of the mitigation measures have trade-offs and limitations. The relationship between the effectiveness of these measures and their trade-offs and limitations is a matter of ongoing balance and refinement. This will be best enhanced through inviting independent scrutiny and assessment. By way of summary and conclusion we note:

- Multi-platform collaborative measures play the greatest role in enabling platforms to rapidly classify and intervene in new objectionable content during an OCC. Conversely, multi-platform collaboration presents risks to human rights and requires measures to enhance auditability and transparency of such collaborations.
- The most effective measures of limiting an OCC are: the GIFCT Content Incident Protocol (CIP) and the GIFCT shared hash database. The greatest limitations of the CIP and the shared hash database are the ways they use automation to remove content rapidly, which creates avenues for potential abuse through collaboration with states, and must remain non-transparent in order to avoid gaming or abuse by perpetrators.
- In situations where a crisis falls short of activating a CIP, the two response protocols developed by GIFCT in partnership with other bodies will play an important role, but it is difficult to assess the effectiveness of these response protocols from an external perspective.
- The platforms must remain focused on enhancing the speed with which they can reliably detect and classify content. For this reason, ongoing improvements in content moderation systems are an essential prerequisite for responding to an OCC. Importantly, the speed with which these systems can classify content should not come at undue cost to the accuracy and reliability of these systems, although some trade-offs are inevitable. Investors could support the improvement of content moderation systems by advocating for measures that enhance transparency and support the development of shared bodies of expertise in how content moderation is conducted at scale.

We are skeptical of any suggestion that changes by the platforms to limit access to livestreaming services will play a strong impact on mitigating future OCCs, given the role of non-livestreamed content in the OCC related to the 15 March terror attacks. Further, the trade-offs of limiting access to this technology are significant.

Key insights for the investor group

What follows are key insights from Brainbox's investigation into the kinds of mechanisms implemented by the platforms to respond to objectionable content crises (OCCs).

- The platforms are content moderation businesses: some commentators go so far as to say that the platforms are not platforms without content moderation, in the sense that content moderation is and always has been an inherent feature of what it means to be "a platform".[47] We have concluded that the platforms' core business activity is amplifying and suppressing different digital content, usually based on the identified preferences and interests of users,[48] the rules and preferences of each platform, as well as the requirements of domestic law.[49] This will continue to be the case whether or not the platforms are ever subject to further State-based regulation.
- There is no prospect that the platforms' content moderation load will decrease, unless we see a large-scale transition away from public social media platforms toward end-to-end encrypted messaging platforms, which would diminish the platforms' user bases as well as limiting providers' ability to moderate content in encrypted channels. So long as the platforms continue to host user-generated content, the requirement to conduct content moderation will continue. Further, to the extent that states begin to impose substantial penalties or other regulatory interventions to influence the way the platforms moderate content, this will only increase the content moderation burden on the companies.
- Content moderation is a complex exercise that is only going to increase in complexity. The platforms' successes in moderating content according to various standards using manual and automated techniques will be a significant influence on their long-term social, political and economic impacts.
- It is important to be critical of decontextualised percentages or numbers offered by the platforms to illustrate how they are conducting content moderation.
    - This includes statistics intended to illustrate the scale of content moderation efforts or the success of those efforts but which withhold any indication of proportion or effectiveness. For example, while the platforms may volunteer statistics around the raw number of hashes in their database, this number is meaningless without further information. The same is true for percentage claims offered by the platforms, where 1% (or even less than 1%) may still be a number measured in thousands or millions.
    - When presented with metrics used in reporting on content moderation, it is important to request clarification and detail, particularly during an attack. For example, Facebook refers to "1.5 million re-uploads" in the days following the Christchurch attacks as an indication of the scale of the challenge they were facing. But this metric could be interpreted in different ways: some

[47] Gillespie, Tarleton (2018) Custodians of the Internet. Yale University Press. Kindle Edition.

[48] Platforms aim to understand users' preferences and interests from tracking their online activity within and outside of the platforms.

[49] Advertising leads are then sold based on the value of privileged access to potential customers. The comparative value offered of advertising through the platforms is superior precision, frequency, familiarity, and quantifiability. Data about user online activity and preferences is also a valuable commodity.

interpretations of this metric reflect well on Facebook, whereas others might not. External observers would benefit from understanding what Facebook means when it posits particular metrics, including what those metrics represent and how they are calculated.

- From a long-term perspective, the greatest threats to the success of the kinds of automated cross-platform content moderation tools used by GIFCT will relate to the transparency, fairness, and de-politicisation of those systems.[50] We support any initiatives that contribute toward demonstrating those systems are transparent, fair, and de-politicised (and we would add, deferent toward human rights principles). The platforms will need to take affirmative steps to generate confidence among regulators and the public in their collaborations in these areas. There is some indication that they are anticipating this need, with significant research funding being made available to independent researchers to study such topics.
- We encourage more proactive consideration of how reporting on content moderation activities might be audited by reliably independent groups — this is reflected in our recommendations in Part 2 (regulation). There is also merit to initiatives that enhance the transparency and independence of existing content moderation auditing mechanisms themselves. Auditing mechanisms are not always 100% reliable. For example, the Electronic Frontier Foundation resigned from GNI when the Snowden revelations demonstrated the true extent of data capture between government and platforms was not identified by those audits.[51]
- Commentators have expressed concern about the absence of challenge or appeal mechanisms against the automated take-down machinery built into the GIFCT shared hash database, and the decision to enter content into that shared hash database. The platforms should implement measures to facilitate better scrutiny of GIFCT mechanisms, to support significant work already underway on this point.[52]
- Be vigilant toward the risks created by complete transparency. Platform measures to respond to OCC conditions are comparable to cybersecurity, where organisations are justifiably secretive about how their systems work because it is widely accepted that any knowledge of protective systems will be used to exploit those systems. The March 15 OCC and terror attacks are clear examples of how knowledge of the technology behind automated detection mechanisms will be leveraged to render them ineffective.
- Long-term, we advise closely watching the operation of the Oversight Board established by Facebook, as well as how it is perceived by the public. The performance of the Oversight Board and public response to it are likely to be the clearest signals of whether specialist content moderation will remain a non-governmental task, or whether it will be absorbed into existing state-based legal and regulatory systems. Some commentators have also noted the Oversight Board will effectively become

[50] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945. See also Douek, E 'The Rise of Content Cartels' [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3572309> accessed 31 March 2021.

[51] Douek, E 'The Rise of Content Cartels' [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3572309> accessed 31 March 2021.

[52] We note a range of transparency measures were announced at the GIFCT Global Summit on 26 July 2021: <https://gifct.org/2021/07/26/globalsummit21/>.

one of the most significant sources of legal judgment on matters of human rights and we note the similarities between the Oversight Board and the social media councils endorsed by the United Nations Special Rapporteur on matters of freedom of expression (discussed in depth in Part 2).

- Over time, it will make more sense for content moderation norms and practices to adhere more closely to international human rights norms and standards, because those norms are already subject to broad commitment among nation states.[53] We therefore endorse any activity to promote the adoption of international human rights norms, particularly when it comes to speech regulation by the platforms. Investors should be aware, however, that to apply these human rights norms frequently requires intensely contextually specific case-by-case determinations to be made in relation to individual pieces of content. A human rights approach also requires users to be provided with procedural rights of review and appeal, even in relatively clear-cut cases, as a matter of procedural fairness.
- The adoption of international human rights norms carries trade-offs: for example, applying such norms may work against the interests of nation states, including relatively democratic states as well as the less-democratic ones. In particular, as we discuss in Part 2, this may mean the platforms are obliged to resist regulatory interventions by states which abuse or undermine human rights, including through opposing some regulatory proposals or refusing to comply with state-backed content take-down requests.
- It is not clear from the Royal Commission's Report whether it spoke directly to the social media companies named in our brief. We have also not seen any indication that the platforms have conducted an audit or learning exercise on how their systems were used in the 15 March terror attacks, although it is extremely likely analyses of this kind have been done.
- The use of automated techniques in content moderation creates risks. While automated techniques will be an essential part of content moderation going forward, the investors and state regulators should not aspire to or advocate for a future state where all content moderation is automated. That would be undesirable for a range of socio-political reasons and would arguably produce more negative impacts than it resolves.

[53] Tworek notes comment by UN Special Rapporteur on the Promotion and protection of the Right to Freedom of Opinion and Expression, David Kaye, and notes that "when companies try to have global terms of service, international human rights law on speech makes sense as a starting point", noting that the Facebook Oversight Board charter explicitly mentions human rights. Heidi Tworek, 'Social Media Councils' (Centre for International Governance Innovation, 28 October 2019) <https://www.cigionline.org/articles/social-media-councils> accessed 14 April 2021.

PART 2: WHAT DOES GOOD REGULATION LOOK LIKE?

Introduction and context

Part 1 of this report assessed changes made by Facebook, YouTube/Google and Twitter (the platforms) to respond to user behaviour during an objectionable content crisis (OCC). The investor group has also sought key insights on the topic of platform regulation, particularly in relation to content moderation. This part covers:

- the trajectory of regulatory development as it pertains to the platforms (e.g. what is currently happening, and where do we expect it to head in the future?);
- an analysis of the effectiveness of this trajectory, including positive or negative features in a range of the specified legal instruments that form part of this trajectory;
- an analysis of which regulatory approach is preferable among a range of approaches.

In this part, we refer to and focus on "content moderation regulation" when it comes to platform regulation. By this, we mean **attempts by nation states (individually or in collaboration)[54] to use legislation (a particular form of regulation) passed by legislative bodies to control the way that social media platforms restrict the flow of content between users via their digital infrastructures**. We have made the decision on pragmatic grounds to exclude the areas of copyright, privacy, artificial intelligence regulation, "honest advertising", and antitrust from our investigation, although all of these will play some role too.

This part provides a primer on what good regulation looks like when it comes to content moderation. Our focus in Part 2 is substantially different from the focus in Part 1, including in the following ways:

DIFFERENT ACTORS

Part 1 focuses on the actions by platforms with respect to behaviour by users, whereas Part 2 is focused on actions by nation states, usually in relation to the platforms. When it comes to taking a human rights approach, this is a fundamental difference in orientation. Throughout Part 2, we refer to nation states by simply using the word "states". Unless the context otherwise indicates, we are not discussing federal states within countries such as the United States or Australia.

DIFFERENT CONTENT

In Part 1, we examined specific content: the livestream video of the Christchurch terror attack on 15 March 2019 and the various versions of it. We took it as granted that there was no reasonable argument for the content's continued spread. When analysing regulatory approaches of online content more generally, our focus has turned away from "black and white" content, to content which falls into much more of a "grey" area. One of the core risks with some regulatory proposals we reviewed is that they require platforms to moderate this "grey" content. In particular, some regulators (notably the United Kingdom) are explicitly setting their sights on limiting content which is lawful, but may be "harmful".

[54] For example, via multilateral bodies like the European Union.

The case for intervening in this type of content is less certain. It also means more complex content moderation assessments are required.

A WIDER PLATFORM ECOSYSTEM RAISES CHALLENGES

When it comes to the issue of regulation more generally, there is a risk that regulation extends to a range of actors beyond the core platforms analysed in Part 1 (Facebook, Google/YouTube and Twitter). It is vital that regulators question the implications of any regulatory proposal for a wider range of actors including website owners, internet service providers and content hosts that may lack the level of resourcing available to the large platforms. These smaller actors are sometimes unintentionally caught within states' wide regulatory framing with limited regard for their continued viability. Even the big platforms are likely to find it difficult to implement some of the regulatory proposals we reviewed, even with their financial resourcing and access to qualified people.

Useful framing questions when considering regulatory impacts on a wider ecosystem of platforms

We have found the following questions useful for thinking through the regulatory impacts on a wide range of platforms/actors. These have been repeatedly raised by commentators when discussing the unintended consequences of regulators' proposals. The questions are:

1) What effect would a given regulatory proposal have on sites like Wikipedia?
2) How is the online service being regulated any different from email or text messaging? Would we expect states to regulate non-platform private communications in this way? Would regulation require platforms to break or avoid encryption protecting private communications?
3) How would a particular regulatory proposal apply to comments left on websites which are otherwise unlike a social media platform, such as comments left on news websites or public reviews of goods and services?
4) How might a particular regulatory proposal be abused by a popularly elected executive government administration which is hostile to human rights, democratic norms, and the rule of law?

This last question in particular is the dominant concern from a human rights approach when it comes to granting states legal powers over the digital platforms.

The regulatory trajectory

In our investigation into regulatory interventions for online content moderation, we identified a trajectory of travel — that is the direction that we see approaches to regulation taking. Our research found that a human rights-centred approach to regulatory interventions is preferable. Regulations that focus on transparency and auditing of content moderation systems are generally more likely to be in line with a human rights approach. There are also considerable risks with taking a regulatory approach that focuses on directing platforms how to moderate specific kinds of content, especially where that content is not well defined, or where there are directions to limit "harmful" content outside of that which is already against the law (e.g. child sexual abuse images). Below is a more detailed overview of this trajectory, along with supporting context. The rest of this document draws out these issues in more detail.

STATES ARE INCREASINGLY LOOKING TO INFLUENCE HOW SOCIAL MEDIA CONTENT IS MODERATED

The activities of the core platforms are beginning to touch upon the interests of states, for example by having real and perceived influences on the conduct of elections and providing people and institutions with communication platforms on matters of national significance (whether for extremists or for public messaging and the correction of misinformation). The platforms' activities also affect the rights and interests of citizens within those states' sovereign jurisdictions, triggering states' obligations of protection and promotion of human rights. These states therefore have a legitimate interest in regulating the platforms, insofar as they can justifiably limit some human rights in order to protect others, so long as they do so in a manner that also complies with human rights norms and principles.

There is a greater trend globally toward the use of legislation (i.e. the use of law) to regulate how the platforms moderate content. It is difficult to extricate regulation that affects content moderation from other areas of legislation, such as antitrust, privacy, the use of AI systems, "honest advertising", election interference, misinformation, and other areas which also influence how content moderation is conducted.

As "content moderation regulation" is an emerging area, it is relatively difficult to say what "good" or "bad" regulation looks like. While the human right to freedom of expression is an old topic, the introduction of digital platforms raises many new issues. Unlike traditional analog media, anyone can now create information which can spread quickly to a huge audience.

There are two key questions about content moderation where consensus is still forming:

1. **How should the platforms be moderating particular kinds of content, particularly in a global context?** This question has a procedural element as well as a substantive element. Specifically, it asks what kinds of content should be impermissible, but it also asks what procedures should be followed by platforms and by states in moderating that content. This makes regulating difficult because we cannot always say clearly what we want platforms to do and how to do it.
2. **What is the proper role of government when it comes to using law to influence how the platforms moderate content produced by users?** This makes assessing the quality and effectiveness of regulation difficult. It means we cannot clearly say whether the law sets an appropriate role for the State in relation to platforms and users.

The uncertainty around the answers to these two questions makes it difficult to say with certainty whether a law will be effective, including: whether it will do what we think it does; what unintended consequences may arise; whether it sets appropriate roles for states, platforms, and users; or whether any outcomes it produces are desirable. These questions are also heavily dependent on the wider regulatory system in any particular jurisdiction, including the relative strength of different constitutional actors and other features of their overall legal-democratic system.

HUMAN RIGHTS PROVIDE A WIDELY SUPPORTED FRAMEWORK THROUGH WHICH TO CONSIDER CONTENT MODERATION

In our research, we found widespread support from a range of organisations and commentators for turning to human rights instruments, principles and jurisprudence to generate greater consensus on the questions we outline above. Human rights, and in particular the UN Declaration of Human Rights, create a universal set of standards which are intended to manage the relationships between the rights of individuals and states (and increasingly, commercial entities). Human rights instruments together set out a widely agreed statement about what can and should be done by states when it comes to balancing the rights of individuals, including both users of platforms and the platforms themselves. There should be ready agreement that regulation which undermines human rights without justification is undesirable. Equally, regulation which requires the companies themselves to undermine human rights is also undesirable.

Human rights jurisprudence sets out ways for states to justifiably limit or balance human rights in order to protect other interests. In brief, to limit the human rights of individuals using law, states must comply with the following principles.[55] We have formed conclusions about the overall regulatory trajectory based on this assessment framework.

1. **Legality –** States must use law if they wish to restrict rights. This law must be clear and certain, and minimise the role of discretion in saying what the law means. The law should not permit states to engage in selective enforcement in order to advance states' own interests. Any limitation imposed by law should also be subject to rights of review and appeal to judicial bodies.
2. **Legitimacy –** States should only limit human rights for legitimate purposes, which may include the protection of other human rights, as provided for in human rights law.
3. **Necessity –** States are required to show a demonstrable connection between the limitation imposed on a human right and necessity of that limitation in order to achieve a particular legitimate goal. In a content moderation context, this means that states must be able to show that it is necessary to limit individual human rights in order to protect another interest: it is not enough to merely assert a connection exists.
4. **Proportionality –** States must only limit a human right to the extent necessary to achieve a legitimate aim. Where a range of potential actions are available, states should adopt the course of action which imposes the least restriction on human rights.

In our assessment, one area requiring more detailed attention from regulators relates to the principle of necessity. There must be a demonstrable connection between the proposed regulatory intervention, and real harm to a legitimate interest protected by human rights instruments. This means that states must be able to persuasively show that the thing they are seeking to limit is causing a tangible adverse outcome of the kind they can legitimately protect. To put it bluntly, states have to show that a particular kind of content is really affecting peoples' human rights: it is not enough to point to a vague connection between content and an adverse outcome. In some cases, a human rights approach means that, perhaps regrettably, we have to tolerate certain kinds of harm because that content serves a greater good. If states wish to justify regulatory interventions, then the best thing they can do is to support empirical work exploring the connection between particular types of content and real-world harm.

[55] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35 Human Rights Council. Thirty-eighth session (18 June−6 July 2018).

Historically, nation states have been the greatest threats to the human rights of individuals. This is still the case globally, including in liberal democracies. For this reason, taking a human rights approach to platform content moderation will mean that states must restrain themselves from passing regulation that threatens human rights without demonstrable justification. In many cases, if the platforms are to demonstrate a genuine commitment to a human rights approach, they will be obliged to resist state action, insist upon proof that content is undermining other human rights interests, and to insist on legal and procedural rights that protect platforms and users and restrain the power of the state.

POTENTIAL REGULATORY APPROACHES

Our conclusion is that the strongest case for regulation from a human rights perspective relates to the area of transparency and auditability of content moderation systems. These regulatory approaches would standardise approaches to transparency reporting on how content is being moderated. Associated measures that enhance transparency come from regulation which creates rights of review and appeal for users against content moderation decisions by the platforms, which as a result require platforms to explain how they made the relevant decision and according to what factors.. We found cautious and appropriate support from human rights bodies for this approach to social media regulation. An example is the EU Digital Services Act.

We cannot endorse content-specific standards in regulation, unless these are linked to content which is already illegal (for example, child sexual abuse imagery or incitement to violence). Using human rights as a standard of assessment, we found significant and widespread opposition from a range of groups to this approach to regulation. We believe this opposition is justified. We also urge caution about heavy handed or punitive approaches imposed on platforms that incentivise content moderation practices that are inconsistent with human rights, unjustifiably rely on automation, or fail to consider perverse incentives.

Because states pose a threat to human rights of privacy, freedom of expression and freedom of association (as well as a range of other human rights), it is worth considering the merits of non-state non-legislative regulatory options. These non-state options do not have to be purely self-regulatory. There are a range of non-state regulatory measures that have been introduced (discussed later in Part 2). It is worth giving these time to operate and mature, especially because they may also illustrate what works well and does not work well when it comes to state-led regulation.

FINAL CONCLUSIONS ON SUMMARY OF THE REGULATORY TRAJECTORY

Our report would not be complete without emphasising the following points:

- The human right to freedom of expression is not the same as the right to “freedom of speech, or freedom of the press” in the first amendment to the American Constitution.
- The human right to freedom of expression is not absolute. It can be limited according to human rights law, including by balancing it with other human rights, but only if such limitations are “lawful”, “proportionate”, “necessary”, and “legitimate” as understood in human rights jurisprudence.
- The human right to freedom of expression enjoys wide support from nation states and non-state bodies via the institutional arrangements flowing from the United

Nations and international law. It has been expressed using relatively specific language and has a large amount of accompanying material which explains what it means and how it should be applied in specific situations, including how it should be balanced against other human rights.

- The human right to freedom of expression is regarded by human rights experts as being one of the most important human rights. To put it bluntly, the right to freedom of expression is regarded as deserving even greater protection than other human rights in situations where rights must be balanced. This is because the ability to freely express opinion and share information is essential in a democratic society for the conduct of free and fair elections and to enable people within a state to draw attention to other breaches of human rights by state and non-state actors.
- The right to freedom of expression has always taken account of the way that states seeking to undermine freedom of expression will target a particular medium, technology or infrastructure used to express or seek information.
- In practice, the platforms already set content moderation standards much more restrictively than nation states can or should do, when it comes to the right to freedom of expression.

In the following sections of the report, we explain some of these summary points in more detail.

Defining content moderation regulation

IT IS DIFFICULT TO EXTRACT CONTENT MODERATION FROM A RANGE OF OTHER REGULATORY AREAS

The issue of platform regulation itself is sprawling and diverse. It is difficult to extract one area (ie, copyright protection or individual privacy) from other areas (ie, antitrust and consolidation of commercial power, multinational taxation).[56] This is a difficulty facing regulators too. We have seen a range of approaches to platform regulation, with some legislation that purports to target narrow types of harmful conduct (i.e., the Abhorrent Violent Material legislation in Australia) and others that cover broad areas (i.e., the Digital Services and Digital Markets proposals in the European Union, and the Online Safety Bill in the UK). Each of the interrelated areas interacting with the narrower topic of content moderation are worthy of specialist investigation and research and we can only cover them so far in the context of this current advice.

As discussed above, this section will focus specifically on content moderation regulation. By this, we mean attempts by nation states (individually or in collaboration)[57] to use legislation (a particular form of regulation) passed by legislative bodies to control the way that social media platforms restrict the flow of content between users via their digital infrastructures. We have made the decision on pragmatic grounds to exclude the areas of copyright, privacy, artificial intelligence regulation, "honest advertising", and antitrust from our investigation.

Case Study: Platform regulation in the European Union

The European Union is an important jurisdiction when it comes to regulating the platforms' conduct. There are a number of regulatory proposals and existing regulations which have some bearing on the platforms' conduct. For practical reasons, we need to exclude the following from our analysis, but we anticipate they will play a significant role on how states regulate content moderation, so it is worth providing a short overview:

- The European Commission has announced a proposal to regulate the use of artificial intelligence (AI) systems. What is meant by an AI system is defined broadly at this stage, which is both a strength of the proposal and a challenge facing any attempt to regulate the use of AI. Initial proposals suggest that AI systems must be assessed based on the level of risk they pose to a range of interests. The level of assessed risk will flow through to different supervisory arrangements imposed upon those systems in response. All content moderation systems will deploy algorithmic systems that are likely to fall within the ambit of this regulatory proposal. The proposal is still at a very early stage and only peripherally relevant, but it illustrates the difficulties of taking an atomistic approach to regulation of content moderation given the range of issues and subjects involved.
- The General Data Protection Regulation in the EU covers user privacy and has had sweeping effects on global use of online platforms, and the way that online platforms process user data. Regulatory proposals that touch upon content moderation

[56] Another area that could be considered is the health and safety conditions facing human content moderators, who must view and classify content that is extremely traumatic to see and hear. See Sarah T Roberts "Behind the Screen" (2019) Yale University Press.

[57] For example, the European Union.

will affect user privacy in two ways: first, they may require platforms to disclose information about users who are posting content that infringes State regulatory standards. Second, content moderation itself (if understood as the serving of content to a user based on their identified preferences from information about them) relies on private information, affecting user privacy. Again, investors should be aware of this link. The UN Special Rapporteur on Freedom of Expression has also linked the human right to privacy to the human right to freedom of expression in writing about State regulation of content moderation by the platforms.

PLATFORM CONTENT MODERATION AFFECTS THE INTERESTS OF STATES AND INDIVIDUALS

Nation states have a legitimate interest in regulating the way that content moderation is conducted on the platforms. States are becoming interested in regulating the activities of the platforms. Content moderation may affect the interests of states in the following ways.

1) The platforms are involved in distributing information which is perceived to affect the interests of nation states and democratically elected leaders. These leaders have personal and political interests in how content is moderated.
2) The platforms host information which is thought to contribute to threats to critical social or technological infrastructure (ie, 5G conspiracy theories leading to attacks on cell towers, undermining confidence in election software, proliferation of radicalising content, hate speech and incitements to violence). States have legitimate interests in protecting the integrity of legal, social, economic and political governance systems. Notably, the platforms all recognise this, and have dedicated teams directed toward the detection of coordinated inauthentic behaviour, which is focused on disinformation campaigns by both foreign nation states and non-government entities, some of whom are commercially oriented.
3) The platforms also are a vector to distribute information that can be harmful to users and to victims. At the black-and-white end of the spectrum this includes objectionable content such as the Christchurch livestream or Child Sexual Abuse Material (CSAM). At the more arguable end, there is content like “coordinated inauthentic behaviour” or speech that is “lawful but harmful”. States have an interest in regulating behaviour that harms their citizens and they have a legitimate interest and obligation in protecting human rights within their sovereign borders.

These factors mean that states are taking increased interest in how the platforms operate and beginning to investigate the extent to which regulation might be useful. A human rights approach in fact requires them to protect citizens’ interests where conduct on the platforms may undermine human rights.

NON-STATE REGULATION THAT IS LESSER THAN LEGISLATION STILL HAS VALUE

Another area of complexity to consider when it comes to shaping how platforms moderate content is that the regulatory tools and potential interventions available are much broader than the use of legislation by nation states and other legislative bodies. For example, human rights bodies, including non-government organisations (NGOs) and the broader “civil society” sector also play a role in the coordinated influence of various actors toward an intended set of outcomes. Regulation is not limited to legislation used by nation states.

Importantly, this means the platforms can also fairly be described as taking independent and collaborative regulatory action toward their own operations, toward other platforms, and towards their users. This dynamic becomes even more complex and important when the concept of “algorithmic regulation”, which is currently an area of scholarly interest, is introduced more broadly. Algorithmic regulation refers to the way that the platforms’ own digital structures exert a regulatory effect on what can or cannot be done.

The key insight is that it would be a mistake to frame this area as being a binary choice between platforms regulating their own activities or introducing new legislative tools.. In reality, the tools available to various actors even in a fully “regulated” space are much broader than just legislation, and the range of actors who can take “regulatory” actions are much broader than just states. In this regard, many of the interventions we identified in Part 1 of this advice can be fairly described as regulatory responses.

Human rights create useful standards for assessing regulatory proposals

HUMAN RIGHTS REGULATE THE RECEIVING AND IMPARTING OF ONLINE EXPRESSION

“Content moderation” is part of the centuries-old contest around the proper boundaries of when people should be allowed to impart and receive opinion, expression, and information. Increasingly, the human right to freedom of expression and its associated jurisprudence is shaping the way that platforms decide when, how and why to intervene. Human rights instruments also provide a widely agreed universal statement of when states should be allowed to limit freedom of expression, including in response to a need to protect other human rights. The human right to freedom of expression therefore provides an anchoring framework for this assessment that enjoys wide international agreement. One other benefit of adopting a human rights approach is that we understand human rights instruments to form a significant foundational component of what it means to take a responsible investment approach.

There are two ways that human rights instruments are relevant to legislation or regulation by states:

1. A legislative regime can itself be inconsistent with human rights, or require platforms (or others subject to the regulatory regime) to undermine human rights in order to comply with the law. Human rights create a basis for the international community (including the UN, other states, civil society, or other non-state actors) to articulate objections to nation state’s regulatory regime if it is inconsistent with a human rights approach.

2. In a situation where legislation has been passed, whether or not it is consistent with human rights, human rights remain relevant to the way that legislation is applied and enforced. As such, whether or not legislation requires the platforms to breach human rights, human rights are still relevant to assessing how far citizens in a state are protected by human rights instruments.

The UN Special Rapporteur for Freedom of Expression states that platforms should be applying human rights law, not domestic laws of states, as the “authoritative global standard” for protecting freedom of expression:[58]

> *Companies should recognize that the authoritative global standard for ensuring freedom of expression on their platforms is human rights law, not the varying laws of States or their own private interests, and they should re-evaluate their content standards accordingly. Human rights law gives companies the tools to articulate and develop policies and processes that respect democratic norms and counter authoritarian demands. This approach begins with rules rooted in rights, continues with rigorous human rights impact assessments for product and policy development, and moves through operations with ongoing assessment, reassessment and meaningful public and civil society consultation. The Guiding Principles on Business and Human Rights, along with industry-specific guidelines developed by civil society, intergovernmental bodies, the Global Network Initiative and others, provide baseline approaches that all Internet companies should adopt.*

[58] Above, A/HRC/38/35 at p 20. This recommendation has been reflected in the charter of the Facebook Oversight Board, which applies human rights law, and not domestic law.

The right to freedom of expression in human rights law is found in article 19 of the Universal Declaration of Human Rights (UNDHR) and article 19 of the International Covenant on Civil and Political Rights (ICCPR). In the UNDHR, the right to freedom of expression is described as follows:

> *"Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."*

It is important to understand that the human right to freedom of expression is not the same as the American constitutional right to "freedom of speech". The former (freedom of expression) is based in human rights instruments which enjoy broad support from a range of nation states via the United Nations and its General Assembly. The latter (freedom of speech) is based on ideas expressed in the Constitution of the United States of America, which applies primarily to government actions (this is referred to as the "state action" doctrine).

By way of broad contrast with the American constitutional approach, human rights law requires states to actively protect freedom of expression, not just to refrain from limiting it. This protective obligation is highly relevant to assessing whether a regulatory proposal by a State is desirable and is a core feature of the Guiding Principles on Business and Human Rights.[59]

> *Human rights law imposes duties on States to ensure enabling environments for freedom of expression and to protect its exercise. The duty to ensure freedom of expression obligates States to promote, inter alia, media diversity and independence and access to information. Additionally, international and regional bodies have urged States to promote universal Internet access. States also have a duty to ensure that private entities do not interfere with the freedoms of opinion and expression. The Guiding Principles on Business and Human Rights, adopted by the Human Rights Council in 2011, emphasize in principle 3 State duties to ensure environments that enable business respect for human rights.*

In the Guiding Principles themselves, States have an obligation "ensure that other laws and policies governing the creation and ongoing operation business enterprises ... do not constrain but enable business respect for human rights."[60] Many of the regulatory proposals we reviewed would undermine the platforms' ability to respect users' human rights.

THE HUMAN RIGHT TO FREEDOM OF EXPRESSION CAN BE LIMITED TO PROTECT OTHER HUMAN RIGHTS

The right to freedom of expression is not the only relevant human right to the way the platforms operate. Other human rights can be relevant generally or arise depending on the content in question. The UN Special Rapporteur has identified the rights to privacy, religious freedom and belief, opinion and expression, assembly and association, and public participation among others.[61] Depending on context, other rights can be engaged. For example, in its decision to indefinitely suspend President Donald Trump from Facebook,

[59] Above, A/HRC/38/35 at at para 6.

[60] Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework (2011) HR/PUB/11/04.

[61] Above, A/HRC/38/35 at at para 5.

Facebook's Oversight Board referred to a range of human rights including the rights to freedom of expression, security of the person, non-discrimination, participation in public affairs, and the right to vote.[62]

While a human rights approach is desirable, there is a risk in adopting "human rights approaches" that all discussion becomes "rights-based". Specifically, a rights-based discussion can obscure the fact that most of the issues being discussed are about trade-offs, not absolutes. The reality is that most if not all rights can be limited according to particular processes designed to balance rights and freedoms.[63] Importantly however, there are constraints on how rights can be balanced against one another. Further, if we hope to have a precise discussion about "balancing" rights and freedoms with other rights, or other considerations, then we need a clear-eyed assessment of what is on each side of the scales to be balanced. This can be difficult when the debate is dominated by broad references to unspecified harms arising from vague classes of content.

While human rights must frequently be balanced where they conflict, it is important that the investors are aware of the importance that human rights experts attach to the right to freedom of expression. In a report by the Danish Institute for Human Rights and the Council on Ethics for the Swedish National Pension Funds, the authors write:[64]

> *The human right to freedom of expression underpins democracy and is essential for the protection of all other human rights and freedoms. It includes the freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.*

The authors also note that states also have obligations to protect citizens from particular kinds of expression. "Hate speech" is not a human rights term, but states have a duty to protect citizens from incitement to discrimination, hostility or violence.[65]

> *International human rights standards do not define "hate speech" as such. However, states have a duty to protect individuals against "national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence". ... States have also enacted laws to prohibit and punish online hate speech. Such legislation must be carefully applied to ensure that, while achieving its primary aim, it does not unduly restrict legitimate expression.*

THE HUMAN RIGHT TO FREEDOM OF EXPRESSION CAN ONLY BE LIMITED ACCORDING TO SPECIFIC CRITERIA

The previous UN Special Rapporteur on the right to freedom of expression, David Kaye, summarises the way that the right to freedom of expression may be limited in ways that are consistent with human rights principles.[66] Per article 19 (3) of the Covenant, state lim-

[62] Case decision 2021-001-FB-FBR, Oversight Board, 6 May 2021 at p 15.

[63] We note that even the first amendment right to freedom of speech in the American Constitution can be limited to some degree. By way of illustration, see The Lawfare Podcast: Content Moderation and the First Amendment for Dummies (11 March 2021) with Prof Genevieve Lakier <https://www.lawfareblog.com/lawfare-podcast-content-moderation-and-first-amendment-dummies>.

[64] See report of the Danish Institute for Human Rights "Tech Giants and Human Rights: Investor Expectations" (2021) <https://www.humanrights.dk/sites/humanrights.dk/files/media/document/Tech%20giants%20and%20human%20rights_2021.pdf> at page 18.

[65] Ibid at p 23.

[66] Above, A/HRC/38/35.

itations on freedom of expression must meet the conditions of legality,[67] necessity and proportionality,[68] and legitimacy.[69]

Notable insights we take from Kaye's summary include:

- While the right to freedom of expression can be limited, it can only be limited for particular legitimate purposes. This can include protecting the rights of others, so long as they meet the criteria below (related to legality, necessity, and proportionality).
- To comply with the principle of legality, legal restrictions on freedom of expression must limit government discretion, and there must be a clear and precise distinction between expression that is lawful and unlawful. Further, any application of the law to limit freedom of expression should lead to a right of review and appeal to a judicial body.
- States must show that there is an actual connection between whatever problem they assert exists and the kind of intervention they are proposing to make. This requires a clear demonstration of what kinds of expression are causing particular kinds of harm, and how the legal proposal at hand will target only that kind of expression, and that targeting this kind of expression will be effective.

Many of the regulatory proposals we have reviewed do not meet these criteria because:

- The law is not precise enough to allow people to reliably distinguish between lawful and unlawful expression and permits too much discretion to the state (or to platforms) to determine whether or not content is lawful or not.
- There is no clear link demonstrated between the presence of some kinds of expression and the particular harms that states allege that expression causes. Some of these harms have also not been clearly linked to other human rights interests to be protected.
- There is an uncertain evidence base for assessing how the proposed limitation on particular kinds of online expression will have a demonstrable effect at mitigating the identified harms.

[67] "Legality. Restrictions must be "provided by law". In particular, they must be adopted by regular legal processes and limit government discretion in a manner that distinguishes between lawful and unlawful expression with "sufficient precision". Secretly adopted restrictions fail this fundamental requirement. The assurance of legality should generally involve the oversight of independent judicial authorities."

[68] "Necessity and proportionality. States must demonstrate that the restriction imposes the least burden on the exercise of the right and actually protects, or is likely to protect, the legitimate State interest at issue. States may not merely assert necessity but must demonstrate it, in the adoption of restrictive legislation and the restriction of specific expression."

[69] "Legitimacy. Any restriction, to be lawful, must protect only those interests enumerated in article 19 (3): the rights or reputations of others, national security or public order, or public health or morals. Restrictions designed to protect the rights of others, for instance, include "human rights as recognized in the Covenant and more generally in international human rights law". Restrictions to protect rights to privacy, life, due process, association and participation in public affairs, to name a few, would be legitimate when demonstrated to meet the tests of legality and necessity. The Human Rights Committee cautions that restrictions to protect "public morals" should not derive "exclusively from a single tradition", seeking to ensure that the restriction reflects principles of non-discrimination and the universality of rights."

ADOPTING A HUMAN RIGHTS APPROACH BUILDS CONSENSUS

Many commentators argue that within the international community, as well as within domestic jurisdictions, there is little social consensus about what the platforms should do in relation to specific content. As such, they argue it is too soon to be committing vague rules to legislation. This concern can be avoided in two ways:

1) By adopting human rights approaches which already enjoy broad support across the international community.
2) By limiting content moderation regulation approaches to content which is already illegal, rather than attempting to create new categories of content that is "harmful but lawful" in the way being attempted by the UK Online Safety Legislation.

Importantly, the investors should be aware that states frequently perceive human rights as a barrier which prevents the state from doing what it wants to do. In this regard, states do not always want to protect their citizens' human rights. If the platforms act in order to maximise the protection of human rights (even in proportionate and limited ways that remove the most unarguably harmful content) that does not mean that states will support the platforms to act in that way.

Given it operates in a global context, the Investor Group should be aware that a particular concern held by human rights groups is that authoritarian states will justify repressive regulatory approaches in their own countries by pointing to similarly repressive regulatory approaches being adopted by democratic nations. Furthermore, some of the legislation being adopted creates the risk that it will have extra-territorial effect: for example, a restriction in Germany may lead to content being restricted outside Germany. This would be undesirable for a range of reasons, not least that citizens of other countries have no democratic input into the laws adopted by other nations.

A NOTE ON INTERMEDIARY LIABILITY

The dominant legal and policy position globally is that platforms are not liable legally for the content posted by users. This area of law and policy is referred to as intermediary liability. Historically, if a person objected to content posted on a platform by a user, their legal remedy was against the user, not the platform. The current regulatory trajectory is beginning to run against this historical position because it makes the platforms liable for the content produced by their users. An important point to consider is how far many of the harms being targeted by these regulatory proposals might instead be dealt with better as a user-to-user issue.

Many of the regulatory proposals related to content moderation are trending toward making platforms liable for content posted by users. This means the principles of intermediary liability are being undermined. Intermediary liability is exemplified by the Communications Decency Act of 1996 includes s 230, referred to as "the twenty six words that created the internet" by scholar Jeff Kosseff.[70] Section 230 states:

> *No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.*

[70] Kosseff, Jeff. The Twenty-Six Words That Created the Internet. Ithaca; London: Cornell University Press, 2019.

The digital tech platforms have traditionally been based in the US, meaning US law has influenced their operations.[71] Contemporary discussion about regulation in the US focuses heavily on the notion that section 230 could be repealed, or that protections conferred on companies by section 230 could be revised or restricted (in practice, this has already occurred through copyright protection legislation). We understand that these proposals have not been given much detail.

The effect of section 230 (as it is commonly referred to among internet scholars) is to provide, through US federal regulation, a shield to the platforms, or any website that hosts content generated by users who are not affiliated with the platforms themselves. It is broadly accepted that the platforms could not exist in their current form without s 230: this is said to be because, without it, the platforms would be treated as endorsing or themselves expressing the information that users submit via their platforms, thereby exposing them to legal and financial risks. Balkin has said that:[72]

> *Section 230 immunity and, to a lesser extent, § 512 [copyright] safe harbors have been among the most important protections of free expression in the United States in the digital age. They have made possible the development of a wide range of telecommunications systems, search engines, platforms, and cloud services without fear of crippling liability. An early version of Google or Facebook might not have survived a series of defamation lawsuits if either had been treated as the publisher of the countless links, blogs, posts, comments, and updates that appear on their facilities.*

There is wide consensus that section 230 is not well understood in public discussions. In particular there has been a concerted campaign by some in US politics to suggest that it imposes a requirement on platforms to provide moderation of content that is, on a party-political basis, neutral, objective or balanced, and only provides such a shield if this requirement is followed. This is incorrect.

Section 230 is a kind of "intermediary liability law". In the European Union, the e-Commerce Directive plays a similar role. Intermediary liability laws not only absolve platforms of liability if they do not moderate content, but also allow them to actually moderate content. Kosseff explains that section 230 is, in part, a response to case law which suggested that a website that removed some content, but permitted other content, could or should be treated as endorsing the content that it allowed to remain on its services. As a result, it was treated as being "the publisher or speaker" of the content it did not remove, and subject to, among other things, action for defamation. Therefore, while section 230 provides a shield for the platforms when they fail to take actions that government and civil society might want them to take, such as removing certain types of "harmful" content, it also enables them to take the actions which are being suggested. As some commentators have put it, without section 230, the platforms would be like other corners of the internet, which are practically unusable because they are dominated by content such as pornography and spam.

[71] Increasingly, states are requiring the platforms to have representatives present within their sovereign jurisdictions, not only in the US. This has created concerns about the personal safety of those representatives in countries seeking to co-opt or control the way that the platforms moderate content: two recent examples of this include India and Turkey.

[72] Balkin JM, 'Old-school/New-school Speech Regulation' (2014) 127 Harvard law review 2296 at 2313.

MANILA PRINCIPLES

As noted above, many countries adopt some form of intermediary liability law, which is essential for any platform that hosts user-generated content. Intermediary liability can also be a crucial tool for platforms to resist interference by governments seeking to control the way that platforms moderate speech by taking enforcement action against them. As a result, civil society organisations have articulated the Manila principles on intermediary liability, which date back to May 2015.[73] These relate to the issue of intermediary liability, but notably, they still emphasise transparency and accountability in any laws or content moderation restriction policies and practices.[74] The Manila principles therefore offer a useful standard for assessing regulatory proposals. They include that:

- Intermediaries should be shielded from liability for third party content.
- Content must not be required to be restricted without an order by a judicial authority.
- Requests for restrictions of content must be clear, be unambiguous, and follow due process.
- Laws and content restriction orders and practices must comply with the tests of necessity and proportionality.
- Laws and content restriction policies and practices must respect due process.
- Transparency and accountability must be built into laws and content restriction policies and practices.

There is a credible and notable list of individuals and institutions that support the Manila principles available online.[75]

[73] A comprehensive background document is available at <https://www.eff.org/files/2015/07/08/manila_principles_background_paper.pdf>.

[74] Supported by Electronic Frontiers Foundation, Associacion Derechose Civiles, Article 19, The Centre for Internet and Society, Ong Derechos Digitalses, Kenya ICT Action Network, OpenNet.

[75] See <https://manilaprinciples.org/organization-signatories.html> and <https://manilaprinciples.org/individual-signatories.html>.

What does good regulation look like?

There is widespread support for anchoring the task of regulating content moderation in the language and law of human rights, and particularly the right to freedom of expression. This context provides a vocabulary and a set of widely accepted concepts and principles for articulating why freedom of expression is important, when it can be limited in ways that are consistent with the public good, and what the common risks are to any situation where states are proposing to limit it.

To illustrate the utility of anchoring this topic in the human right to freedom of expression, below we have summarised key points made by the UN Special Rapporteur on Freedom of Expression, David Kaye, in his report to the UN Human Rights Council of 2018.[76] Kaye is a respected academic who has engaged extensively with governments, platforms, academia and human rights groups in formulating his position. He also regularly defends his approach to human rights and the platforms and welcomes engagement with sceptics and dissenting voices.[77]

SUMMARY

In general, good regulation will be clear enough and specific enough that the platforms are able to interpret it and comply with it without recourse to litigation. Bad regulation will leave it unclear what the platforms are required to do, and in relation to what specific content.

The nature of the content is integral to this. If regulation only offers vague indications of what content the platforms should moderate then compliance will become a fraught activity, likely with adverse outcomes for users. This is true even if the regulation is clear in other regards. For example, if the regulation applies to child sexual abuse material, it will be easy to understand what content is relevant and how it should be moderated. If it applies to vague and contestable classes of information – like “coordinated inauthentic behaviour”, “misinformation”, and to a lesser (but still considerable) extent “hate speech” and “terrorist content”, then the platforms have insufficient guidance about what they ought to be moderating and why. With heavy penalties at stake, they will be obliged to err on the side of over-removal of content, which will likely put them in breach of human rights.

As well as applying to clearly defined content, good regulation will create obligations on the platforms to record and report on how they are complying with the legislation. This will complement regulatory requirements for systems of appeal which meet the principles of natural justice, so that users can contest the content moderation actions of the platforms.

At the same time, good regulation will be alert to the unprecedented risks that arise from the platforms attempting to comply. The platforms are an exceptionally advanced architecture for surveillance and create significant digital tools for enhancing or suppressing rights of expression, opinion, privacy, and association. To the same extent that the platforms at their best greatly enable these rights, at their worst, they could greatly repress them.

[76] Above, A/HRC/38/35.

[77] ‘The Lawfare Podcast: The Arrival of International Human Rights Law in Content Moderation’ (Lawfare, 27 May 2021) <https://www.lawfareblog.com/lawfare-podcast-arrival-international-human-rights-law-content-moderation>.

Anjum Rahman is a leading voice on New Zealand's response to the Christchurch attacks. She is a lead spokesperson for Inclusive Aotearoa Collective Tahono, a spokesperson for the Islamic Women's Council of New Zealand, and a civil society member for the panel advising the GIFCT. She wrote an opinion piece for the Guardian opposing elements of New Zealand's own regulatory response to the Christchurch attacks. Within her criticism, she offers a useful summary of a key risk that good regulation will guard against:[78]

> *Any legislation should be assessed considering the worst-case scenario … How might a hostile government misuse this legislation, and what checks and balances are in place to prevent that misuse?*

The key features to look for in assessing regulatory proposals are as follows:[79]

- **Legality:** does the regulation create a clear enough distinction between permissible and impermissible content, such that it avoids abusive or discriminatory enforcement? Can users and platforms reasonably understand what content is or is not allowed, and what they must do in response?
- **Necessity, legitimacy and proportionality:** does the regulation aim to achieve a legitimate aim, such as protection of the rights of others? Is there a demonstrable connection between its purpose and the tools it uses to achieve that aim? Does it take the least restrictive approach available for individual rights and freedoms?
- **Transparency:** does the legislation require records to be kept showing how the legislation has been applied, both by platforms and by governments? Transparency is a crucial measure for detecting and remedying abuse, as well as assessing whether interventions are necessary and proportionate, including whether they are having the intended impact.
- **Reliance on automation:** does the regulation effectively require the use of automated tools in order to achieve compliance? For example, is it impossible to comply with a particular timeframe without the use of proactive algorithmic detection?

In summary, good legislation will do the following:

- Set rules and standards that are as clear as possible to distinguish between what kind of content is permitted and what kind of content is not permitted. "[I]t is not enough that restrictions on freedom of expression are formally enacted as domestic laws or regulations. Instead, restrictions must also be sufficiently clear, accessible and predictable (CCPR/C/GC/34)."[80] Any vagueness or ambiguity in these content standards create a risk that they will be enforced in a discretionary or discriminatory way. This could lead to abuse by nation states against minority populations or political opposition. It could also lead to discriminatory enforcement against minority groups as a result of systemic discrimination, racism, sexism, or other factors.
- Any decision that applies the law must be capable of review and appeal by a legal body, such as a court or tribunal. "While it is recognized that business enterprises

[78] 'Livestreaming Bill Introduced after Christchurch Attacks Could Criminalise Innocent People | Anjum Rahman | The Guardian' <https://www.theguardian.com/world/commentisfree/2021/mar/15/livestreaming-bill-introduced-after-christchurch-attacks-could-criminalise-innocent-people>.

[79] Summarised by Brainbox based on Kaye, above, A/HRC/38/35.

[80] Comment of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression to the Government of Germany on the the draft law "Netzdurchführungsgesetz"(NetzDG) Reference: OL DEU 1/2017 (1 June 2017).

also have a responsibility to respect human rights, censorship measures should not be delegated to private entities (A/HRC/17/31). States should not require the private sector to take steps that unnecessarily or disproportionately interfere with freedom of expression, whether through laws, policies or extralegal means (A/HRC/32/38)."[81]

- There must be a demonstrable connection between the kind of conduct being restricted by regulation and the kind of harm that is alleged to result. Regulation must only intrude upon individual freedoms to the extent necessary to achieve the desired effect. Regulatory aims must also be legitimate in terms of human rights norms. These requirements are caught by the terms "necessity", "proportionality", and "legitimacy": "The requirement of necessity … implies an assessment of the proportionality of restrictions, with the aim of ensuring that restrictions "target a specific objective and do not unduly intrude upon the rights of targeted persons". The ensuing interference with third parties' rights must also be limited and justified in the interest supported by the intrusion (A/HRC/29/32). … [T]he restrictions must be "the least intrusive instrument among those which might achieve the desired result" (CCPR/C/GC/34)."[82]
- Legislation should require and foster transparency about what content moderation actions are being taken and why. These transparency requirements should be imposed on both states and platforms.
- Legislation that requires the use of automated detection and enforcement tools will require companies to use tools that import bias, and may have discriminatory effects. Automated tools are not sophisticated enough to include assessment of context, which is frequently essential for determining whether content is permissible or impermissible.
- Legislation should not unjustifiably limit individual privacy, including by requiring platforms to report users to governments based upon what they are saying or doing online. The right to privacy and the right to freedom of expression are linked.
- Legislation that imposes massive financial penalties will influence platforms to take a course of action most likely to reduce their own risk, including to over-remove content rather than risk a fine, which is likely to have disproportionate or discriminatory effects.

Kaye, as UN Special Rapporteur for freedom of expression, made the following recommendations for States:[83]

- "States should repeal any law that criminalizes or unduly restricts expression, online or offline."
- "Smart regulation, not heavy-handed viewpoint-based regulation, should be the norm, focused on ensuring company transparency and remediation to enable the public to make choices about how and whether to engage in online forums. States should only seek to restrict content pursuant to an order by an independent and impartial judicial authority, and in accordance with due process and standards of legality, necessity and legitimacy. States should refrain from imposing disproportionate sanctions, whether heavy fines or imprisonment, on Internet intermediaries, given their significant chilling effect on freedom of expression."

[81] Ibid at p 2.

[82] Ibid at p 2.

[83] Above, A/HRC/38/35 at paras 65-69, pp 19-20.

- "States and intergovernmental organizations should refrain from establishing laws or arrangements that would require the "proactive" monitoring or filtering of content, which is both inconsistent with the right to privacy and likely to amount to pre-publication censorship."
- "States should refrain from adopting models of regulation where government agencies, rather than judicial authorities, become the arbiters of lawful expression. They should avoid delegating responsibility to companies as adjudicators of content, which empowers corporate judgment over human rights values to the detriment of users."
- "States should publish detailed transparency reports on all content-related requests issued to intermediaries and involve genuine public input in all regulatory considerations."

THE BENEFITS OF TRANSPARENCY REPORTING

A consistent thread throughout the material we have reviewed, including from Kaye, is the role that transparency reporting and transparent approaches can play as a crucial safeguard against abuse by platforms, users or governments. Kaye points to companies' transparency reports as being a key indicator of the kind of pressure being placed on platforms by nation states, but notes that enhanced transparency requirements would produce more useful information:[84]

> *Companies have developed transparency reports that publish aggregated data on government requests for content removal and user data. Such reporting demonstrates the kinds of pressures the companies face. Transparency reporting identifies, country by country, the number of legal removal requests, the number of requests where some action was taken or content restricted and, increasingly, descriptions and examples of selected legal bases. However, as the leading review of Internet transparency concludes, companies disclose "the least amount of information about how private rules and mechanisms for self-and co-regulation are formulated and carried out". In particular, disclosure concerning actions taken pursuant to private removal requests under terms of service is "incredibly low".*

Kaye's recommendations also support platforms to push back against requests by governments to deal with content in particular ways:[85]

> *Companies often claim to take human rights seriously. But it is not enough for companies to undertake such commitments internally and provide ad hoc assurances to the public when controversies arise. Companies should also, at the highest levels of leadership, adopt and then publicly disclose specific policies that "direct all business units, including local subsidiaries, to resolve any legal ambiguity in favour of respect for freedom of expression, privacy, and other human rights". Policies and procedures that interpret and implement government demands to narrow and "ensure the least restriction on content" should flow from these commitments. Companies should ensure that requests are in writing, cite specific and valid legal bases for restrictions and are issued by a valid government authority in an appropriate format. When faced with problematic requests, companies should seek clarification or modification; solicit the assistance of civil society, peer companies, relevant government authorities, international and regional bodies and other stakeholders; and*

[84] Above, A/HRC/38/35 at p 13.

[85] Above, A/HRC/38/35 at para 50.

> *explore all legal options for challenge. When companies receive requests from States under their terms of service or through other extralegal means, they should route these requests through legal compliance processes and assess the validity of such requests under relevant local laws and human rights standards.*

Kaye notes that transparency practices are essential for illuminating the relationships between platforms and states, for supporting trust and confidence, and for mitigating the potential for abuse:[86]

> *In the face of censorship and associated human rights risks, users can only make informed decisions about whether and how to engage on social media if interactions between companies and States are meaningfully transparent. Best practices on how to provide such transparency should be developed. Company reporting about State requests should be supplemented with granular data concerning the types of requests received (e.g., defamation, hate speech, terrorism-related content) and actions taken (e.g., partial or full removal, country-specific or global removal, account suspension, removal granted under terms of service). Companies should also provide specific examples as often as possible. Transparency reporting should extend to government demands under company terms of service and must also account for public-private initiatives to restrict content, such as the European Union Code of Conduct on countering illegal hate speech online, governmental initiatives such as Internet referral units and bilateral understandings such as those reported between YouTube and Pakistan and Facebook and Israel. Companies should preserve records of requests made under these initiatives and communications between the company and the requester and explore arrangements to submit copies of such requests to a third-party repository.*

Transparency is also an important safeguard against the misuse or over-use on automation and the ensuing impact on freedom of expression:[87]

> *Notwithstanding advances in aggregate transparency of government removal requests, terms of service actions are largely unreported. Companies do not publish data on the volume and type of private requests they receive under these terms, let alone rates of compliance. Companies should develop transparency initiatives that explain the impact of automation, human moderation and user or trusted flagging on terms of service actions. While a few companies are beginning to provide some information about these actions, the industry should be moving to provide more detail about specific and representative cases and significant developments in the interpretation and enforcement of their policies.*

GOOD LEGISLATION INCLUDES LEGAL AND CONSTITUTIONAL SAFEGUARDS, WHICH CAN GENERATE UNCERTAINTY

Once legislation passes through democratic processes founded on the rule of law and human rights, its ultimate effect can be different than originally intended. In relation to each state-based regulatory proposal, there are a number of legal and political factors that may lead legislation as drafted to take a different course than expected. This is especially important to bear in mind when it comes to understanding what a regulatory proposal would do based on statements by media or politicians. It is equally important even where regulation has been passed as legislation by democratic bodies. When it comes to

[86] Above, A/HRC/38/35 at para 52.

[87] Above, A/HRC/38/35 at para 62.

protecting fundamental rights, it is a good thing that the intended effect of the law may change over time as it makes its way through political, legislative and legal systems.

When assessing the merits of any regulatory proposal we emphasise the following insights about the legislative process to the investor group.

- Draft pieces of regulation may be subject to significant revision after public consultation, which can take significant time and lead to large volumes of material and commentary. Much of this commentary must be incorporated into a draft in some form. A proper public consultation process should be open to the possibility that the legislative exercise is abandoned entirely, although this is unlikely.
- Even a well-advanced legal proposal from the executive branch of government may not receive enough votes in a legislative body to be passed. This can result from political or pragmatic factors unrelated to the merits of the regulatory proposal.
- Given the long timeframe for developing these regulatory proposals and passing them into law, executive governments pushing a particular piece of legislation may lose power in democratic elections before the regulatory proposal can be advanced. Again, this may have little to do with the merits of the regulation being proposed by that government.
- Where regulation is passed as legislation, within some constitutional systems, the law may be struck down by the judicial branch of government for non-compliance with higher law such as a written constitution. This can render a regulatory proposal ineffective when it eventually comes to be assessed by the judicial branch of government, but it can take a long time before the legal position is clear.
- In some constitutional systems, the judicial branch may interpret statutory language broadly or narrowly in order to take a more rights-consistent approach. While this does not amount to formally striking down a law, it can significantly alter the actual effect of the regulation by comparison with the publicly stated intent of executive or legislative government. Many of the regulatory proposals we investigated used vague or imprecise standards to delineate between permissible and impermissible content, and we predict that judicial interpretation will be a significant factor in understanding what these standards actually require.
- When a law has been passed, compliance can be achieved in two ways: through proactive response by the regulated entity, either for non-punitive good faith reasons, or to avoid punitive penalties; or by compliance and enforcement action by the specified regulatory body. Where regulation relies heavily on punitive deterrent approaches (as many of the proposals do), compliance action will occur through legal processes and may be challenged through review and appeal to courts. As a result, any remedy or penalty imposed on the platforms under the law may take years to be implemented, making it difficult to assess the merits of the regulation in advance. Further, during legal enforcement processes, the practice of statutory interpretation by the judicial branch may substantially alter the effect of the law by comparison with what observers or executive/legislative actors thought the law's effect would be.
- Many of the legislative proposals require further regulatory processes to be followed, for example the articulation of secondary or delegated legislation and codes of practice to narrow the broad drafting of the regulatory instrument as drafted. For this reason, it is impossible to say at this point what the platforms' final legal obligations will be with regard to content moderation.

- In some cases, domestic regulatory proposals may be subject to review by international legal bodies. This is another way that a domestic regulatory proposal as drafted may be challenged or subject to review.

Given the novelty of each of the legislative proposals we assessed, it is difficult to predict how the stated intention of executive or legislative branches of government about a legislative proposal's intended effect will come to pass, given the input of other democratic and legal institutions.

It is important to emphasise that each of the points raised in the bullet points above are important features of proper governance according to the rule of law, the doctrine of the separation of powers and democratic principles. The factors we identify in the bullet points are not barriers to effective regulation, they are part of the legal environment that effective regulation should anticipate and embrace. Such barriers are an essential part of the requirement that states only limit the human right to freedom of expression using law, rather than arbitrary discretion. They result from legal and constitutional efforts to prevent the concentration of power in any one arm of government, precisely because of the threats this poses to individual civil liberties and human rights. Because of the inevitable impact that state influence on platform content moderation will have for rights to privacy, freedom of expression, and freedom of association, it is essential that good regulation embraces and invites input from a range of constitutional actors.

Risks created by content-specific regulatory proposals

We reviewed secondary materials and commentary around a range of regulatory proposals. Our review could not be comprehensive given the limited scale of this project, however we have identified a range of prominent features which are present to varying degrees in a range of regulatory proposals. We set these out here with a view to providing insight into the desirability of those regulatory features.

TREND TOWARD USE OF STATE-LEVEL LEGISLATION AND AWAY FROM SELF- OR CO-REGULATION

There is a clear trend towards greater regulation of the platforms. Much of this is already in effect via a network of self-regulatory and co-regulatory mechanisms and policies, private and independent regulatory institutions, and voluntary collaborations between the platforms and states. There are significant gaps in this network of regulatory mechanisms, specifically in relation to record keeping, public data availability, transparency, and consistency of content moderation decision-making. This variety of non-legislative regulatory interventions is discussed by the Global Network Initiative:[88]

> *Governments are increasingly considering ways to regulate content and conduct through tried-and-true legal demands to intermediaries, to deploying government-ordered network disruptions. Governments are also trying out "new school," less-direct, and non-legal approaches, including pressuring ICT intermediaries to expand the range of content prohibited under their community standards, as well as their enforcement of those standards — often under the (implicit or explicit) threat of legislation or regulation.*

CONTENT-SPECIFIC REGULATION GENERALLY INCORPORATES THE FOLLOWING PATTERN

All content-specific content moderation regulation requires platforms to do a series of common tasks. We covered many of those tasks in Part 1. If legislators wish to commit these tasks to legislation, they must be capable of clearly stating what those tasks are and how they should be performed, which is difficult.[89] Further, each of these tasks imports a degree of risk from a human rights perspective. We explain those core common tasks and illustrate how they import human rights risks below. The tasks are:

1. The legislation must define a set of categories of content that clearly distinguish between permissible and impermissible content. This creates risks to freedom of expression because it authorises states to dictate what may or may not be said. While states clearly can justifiably limit freedom of expression, it is difficult to do so using clear language that creates predictable categories. Traditionally, these categories were framed in terms of legal and illegal speech: increasingly, regulators are attempting to create new categories of speech that are lawful, but allegedly harmful. This is a difficult exercise. It also creates risks that states will unjustifiably limit freedom of expression on particular topics in the name of "safety" or avoiding "harm",

[88] See Global Network Initiative "Content Regulation and Human Rights: Analysis and Recommendations" (2020) at p 7.

[89] This is essential from a human rights perspective, as the line between permissible and impermissible content must be clearly stated as a matter of law, not as a matter of discretion, which might be abused by nation states.

without meeting the thresholds of necessity and proportionality required by human rights law.

2. Next, platforms must be capable of identifying content which allegedly falls into a category requiring it to be dealt with in a particular way. This is a difficult technical task. It is also a difficult legal task, as categories of content may not be well-defined. The current technical methods for identifying such content all have risks of inaccuracy, and all raise the risk that they are abused. Because technical methods are algorithmic, this also raises issues of bias and ethics which are acquiring prominence in discussions in policy and tech communities. Accordingly, while automated methods might be thought to decrease human input, conversely, the use of automated techniques can serve to compound the level of human input and oversight required. The platforms broadly have the following methods available for identifying content which might fall into prohibited categories and each of these methods also raise human rights risks.

    1. **Proactive algorithmic detection (before upload or distribution).** The use of "upload filters" is one of the most restrictive forms of restriction on freedom of expression. They also rely on algorithmic methods which can suffer from unjustified bias, discriminatory effects, or false positives. These sorts of upload filters are the kind used by some companies to implement the GIFCT shared hash database and they are, or should be, reserved only for the most extreme content which has already been appropriately classified.

    2. **Reactive algorithmic detection (after upload and distribution).** Algorithmic detection of content that occurs once something has already entered a platform's systems more likely relies on the use of machine learning techniques and artificial intelligence, which increases the prospect that algorithmic false positives or false negatives will have discriminatory or harmful effects. Algorithmic systems are poor at detecting the context for content, which can drastically change its intended and received meaning.

    3. **Flagging by general users**: users on the platform can flag content that they allege breaches community guidelines or other relevant rules, including regulation. These mechanisms are often subject to abuse or misreporting, for an extremely broad range of reasons which are difficult to anticipate. Such abuses include harmful reporting without justification in order to interfere with another user's online behaviour, including the practice of "copyright trolling". This has led to regimes of "trusted flaggers", as outlined below.

    4. **Flagging by trusted users:** platforms frequently confer "trusted" status on some users, which enables their reports about content infringing platform terms of service to be escalated rapidly. Such regimes create broad issues for who can become a trusted flagger and can also lead to over-reliance on such mechanisms, in situations where flagging by "non-trusted users" is equally important. There is evidence that the platforms are conferring trusted flagger status on some government agencies, which creates further freedom of expression issues.

    5. **Flagging by states:** there is a growing trend where states use platforms' terms of service in order to have platforms take down content. Content is removed on the basis that it infringes platform terms of service, but this obscures the fact that the request has come from a state, and may not have been authorised by any other legal instrument. This creates issues where the user subject

to the takedown notice has no knowledge that a state is the actor responsible for asking for their content to be removed. Further, because platforms are under significant pressure from states on multiple fronts (including in the areas of taxation, antitrust, and content moderation regulation proposals) platforms have a real or perceived interest in being unduly deferential toward notifications made by states. Human rights advocates express concern about the way that platforms do not report on the number of requests for takedown pursuant to community guidelines that are made by states. This practice has been challenged in and upheld by the Israeli Supreme Court, and examined by the UN Special Rapporteur for Freedom of Expression.

3. Once potentially infringing content has been identified, platforms must assess it and classify it into one of the categories created by regulation. At this point, platforms are likely to be confronted with a range of conflicting rule sets, or ambiguities in those rule-sets. For example, platforms must already adjudicate between domestic law, their own content policies, and international human rights principles. In the future, one risk is that platforms will have to be able to apply a range of distinct state-based regulatory mechanisms, all of which have similar structural flaws.

4. Having classified the content once it has been identified, platforms must act in accordance with that classification decision. For platforms, how they should act in response is not always clear cut. This is compounded by the way that platforms sit across jurisdictions. One concern raised by human rights organisations is that domestic regulation may be applied to limit people from viewing content outside the jurisdiction of a state's sovereign jurisdiction. Furthermore, there is a developing discussion about moving content moderation action beyond the binary decision to take it down or leave it up, including to place content behind an interstitial barrier, algorithmically de-rank it so it is seen less often, to delete it, or make it less likely to be seen by users with particular characteristics (ie, children or people expressing a preference not to see content of that type). One risk created by regulation that imposes harsh penalties for non-compliance and sets short timeframes for action is that these decisions must be made fast, and are usually done with a view to avoiding liability, rather than protecting the rights of users, or with regard to the potential abuses by states or other users.

5. Regulation seldom focuses on providing procedural rights or a right of appeal against a decision made by a platform to act on particular content, including whether the platforms' own classification decision was correct. Good regulation should take these process rights into account (the EU's Digital Services Act proposal is the strongest in this regard). One significant criticism of Germany's NetzDG law is that it requires platforms to make classification decisions based on legal criteria, but it provides little judicial oversight or appeal to judicial mechanisms in response to those decisions. One effect of this is that it effectively shifts the state's legal influence over content moderation outside the oversight of the judicial system, removing key checks and balances on government use of legal power.

Preferable regulatory approaches

Among human rights commentators, there seems to be a consensus that regulatory approaches which dictate to the platforms particular content standards are undesirable and create significant risk to human rights. By contrast, there is widespread support from a range of actors for regulatory approaches which enhance independent insight into how the platforms are moderating content according to their own standards.

IMPOSE TRANSPARENCY, DISCLOSURE AND AUDITING FRAMEWORKS FOR PLATFORM CONTENT MODERATION

The best kind of regulation by states would enhance auditing and scrutiny of how content moderation systems are operating. Regimes of this kind are beneficial because of the way they aim to remedy power imbalances between different parties to a content moderation process. We think that the separation of powers, and the use of checks and balances on different actors within content moderation systems, are useful theoretical traditions to draw upon in regulating the content moderation relationship.

This transparency-centred approach is being suggested for the EU's Digital Services Act and it merits further investigation. A similar approach to regulation (in the context of disinformation) is supported by the UN Special Rapporteur on Freedom of Expression:[90]

> *State regulation of social media should focus on enforcing transparency, due process rights for users and due diligence on human rights by companies, and on ensuring that the independence and remit of the regulators are clearly defined, guaranteed and limited by law.*

We endorse regulation of this kind because we believe it contributes to an overall balance and separation of powers between platforms, users and states in the following ways:

- Empower individuals against platforms: to better understand and demonstrate how the platforms may be acting to control their behaviour in the digital space. This will also facilitate better public discussions about how the platforms are acting in relation to particular pieces of content.
- Empower NGOs against platforms by providing a superior evidence base for praising or criticising the way the platforms are moderating content. NGOs will also have better insight into how governments may be influencing platforms.
- Empower states against platforms by providing a better empirical basis for demonstrating where content moderation decisions are being made according to improper considerations or poor process, or where decisions are out of step with community expectations. States would also be able to demonstrate that a particular incident that has attracted public scrutiny is or is not part of a broader pattern of conduct rather than being an isolated incident.
- Empower platforms against states: platforms retain the authority to apply their own content moderation policies, including to apply international standards rather than domestic standards, and to develop an evidence base to defend their content

[90] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan "Disinformation and freedom of opinion and expression" Human Rights Council, Forty-seventh session, 21 June–9 July 2021 (A/HRC/47/25) at para 91.

moderation practices. Equally, a range of competitors will be subject to the same compliance and reporting obligations, meaning that market participants are not penalised for adopting virtuous reporting practices by comparison with their competitors.

- Empower platforms to resist unjustified allegations made by individuals, NGOs or States. Some commentators have pointed to the way that content moderation editorial stories or lapses in content moderation systems are a convenient source of news for journalists. Equally, platform CEOs are being compelled to attend investigative hearings by politicians and being asked about individual content moderation decisions and whether they will take unilateral executive action in response, when in many cases, it is preferable to leave such decisions to appropriate decision-making processes. Equally, non-government organisations, researchers and academics must currently speculate about what the platforms are or are not doing: transparent auditable reporting would have the effect of providing an evidence base which supports the concerns being raised or does not.
- Importantly however, regulation of this minimises the risk that States are empowered against individuals, so long as privacy-preserving practices are adopted. Freedom of expression is a grey area, whereas privacy is an area of law and technological practice that is very well-examined across a range of subject areas. Reporting at a system level ought to be able to be done, in most cases, in a manner that preserves individual privacy.

Notably, transparency and reporting regimes are an essential enforcement mechanism for some human rights instruments. Passing regulation that implements such regimes for content moderation by platforms are therefore reflective of modern human rights practice. Recent human rights instruments such as the United Nations Convention on the Rights of Persons with Disabilities, for example, create obligations on states to collect statistical data that illustrates the extent of their own compliance with the Convention.[91] The Committee on the Rights of Persons with Disabilities also invites "shadow reports" from NGOs and disabled people to enable independent criticism of states' self-reports on their compliance. An important reason for obliging states to collect such data is to empower individuals with disabilities to use that data to illustrate how states are non-compliant with human rights instruments in domestic and international fora.

In our opinion, the best regulatory approach is to standardise the metrics for content moderation, and requires those metrics to be made public. Obligations should also be put on platforms to report on their content moderation efforts in line with these standardised metrics.[92] This will generate a basis for future regulatory action, if required. This evidence base will also assist to take discussions about regulation from anecdotal case-based insights, to broader evidence-based insights at a system level. There remains a risk that reporting on content moderation may raise issues for individual privacy, but these are well known issues that can be navigated, managed, or largely avoided with the right approach.

[91] United Nations Convention on the Rights of Persons with Disabilities (3 May 2008) Article 31. "States Parties undertake to collect appropriate information, including statistical and research data, to enable them to formulate and implement policies to give effect to the present Convention. … The information collected in accordance with this article shall be disaggregated, as appropriate, and used to help assess the implementation of States Parties' obligations under the present Convention …".

[92] There are some indications that this approach is being pursued in regulatory approaches, although at this stage in relation to ad-targeting only: 'Lawmakers Want to Force Big Tech to Give Researchers More Data' (Protocol — The people, power and politics of tech, 20 May 2021) <https://www.protocol.com/policy/social-media-data-act>.

SANTA CLARA PRINCIPLES

If the investor group is minded to advocate for regulation which imposes transparency reporting obligations, then the Santa Clara principles may provide some specificity to what can be broad and non-specific calls for greater transparency.

In an effort to influence the trajectory of state-based and international law regulatory initiatives, groups of interested parties sometimes collaborate on statements of principle about a particular issue. These bodies of principle are not binding on anyone, but they do form influential foundations for regulatory discussions, and points of agreement or departure between interested organisations. In 2018, a group of institutions and individuals with authority in this area met to discuss content moderation at scale. They produced three principles (the Santa Clara principles) on how content moderation should occur.

1. **Numbers.** "Companies should publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines."
2. **Notice.** "Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension."
3. **Appeal.** "Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension."

Such statements of principle can influence regulation by being taken into account by judicial bodies, political bodies, or multilateral institutions such as the United Nations.[93] The statements of principle gain authority when authoritative people or institutions support them – relevant signatories here include the following:

- The American Civil Liberties Union Foundation of Northern California
- The Center for Democracy and Technology
- The Electronic Frontier Foundation
- New America's Open Technology Institute
- Four influential academic writers (Irina Raicu, Nicolas Suzor, Saray Myers West and Sarah T Roberts). Notably, Suzor is now a member of the Oversight Board set up by Facebook.

The Santa Clara principles provide a useful illustration of the kind of information that should be disclosed for content moderation to be adequately transparent. The principles say the following minimum information should be disclosed "in a regular report, ideally quarterly, in an openly licensed, machine-readable format":

- Total number of discrete posts and accounts flagged.
- Total number of discrete posts removed and accounts suspended.
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by category of rule violated.
- Number of discrete posts and accounts flagged, and number of discrete posts

[93] The UN Special Rapporteur for freedom of expression has recommended, for example, that companies adopt "industry-specific guidelines developed by civil society ... and others" as a baseline approach for protecting freedom of expression. Above, A/HRC/38/35 at para 70.

removed and accounts suspended, by format of content at issue (e.g., text, audio, image, video, live stream).

- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by source of flag (e.g., governments, trusted flaggers, users, different types of automated detection).
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by locations of flaggers and impacted users (where apparent).

The principles are a useful starting point for understanding the kinds of information that civil society actors expect the platforms to disclose in order to be transparent.

Non-state regulatory measures to consider

As discussed, because of the risks that state use of legislation may pose to human rights, it is worth considering whether non-legislative regulatory measures that are less subject to control by states are preferable. We provide some examples below of the way that non-legislative regulatory measures can be useful.

PLATFORMS REGULATE THEIR SERVICES THROUGH CONTENT MODERATION

The services provided by large platforms are highly regulated information environments. Within these, platforms are constantly controlling what opinion and information their users are allowed to share and receive, even while they may give their users the impression that they do not. This is a first order form of regulation on content. They do this through a combination of written rules and policies, algorithms, and enforcement actions. The major internet platforms tend to have highly nuanced and comprehensive rulesets about what content may be uploaded, shared, and accessed on their services – though there is variation in these rules across the platforms and there is some concern about the consistency and reliability of the processes followed to apply these content rules.

As such, there is no question that content on the major social media platforms is already highly regulated. As a matter of fact, content moderation is their core business activity. This is best summarised by Gillespie as follows:[94]

> *… platforms do, and must, moderate the content and activity of users, using some logistics of detection, review, and enforcement. Moderation is not an ancillary aspect of what platforms do. It is essential, constitutional, definitional. Not only can platforms not survive without moderation, they are not platforms without it. Moderation is there from the beginning, and always; yet it must be largely disavowed, hidden, in part to maintain the illusion of an open platform and in part to avoid legal and cultural responsibility. Platforms face what may be an irreconcilable contradiction: they are represented as mere conduits and they are premised on making choices for what users see and say. Looking at moderation in this way should shift our view of what social media platforms really do: from transmitting what we post, to constituting what we see. There is no position of impartiality. Platform moderators pick and choose all the time, in all sorts of ways. Excluding porn or threats or violence or terrorism is just one way platforms constitute the social media product they*

[94] Gillespie, Tarleton. Custodians of the Internet. Yale University Press. Kindle Edition.

> *are generating for the audience. The persistent belief that platforms are open, impartial, and unregulated is an odd one, considering that everything on a platform is designed and orchestrated.*

THE CHRISTCHURCH CALL AND EFFECTIVE MEASURES IDENTIFIED IN PART 1

The Christchurch Call is a form of regulatory response which does not rely on the coercive force of legislation at the nation-state level. A foundational principle of the Christchurch Call is that it is voluntary. The Christchurch Call exists at a diplomatic level between nation states and technology companies. It is effectively a statement of principle with a comparable status to other statements of principle, such as the Manila and Santa Clara principles (discussed below), although enjoying greater weight because of the participation of nation states and technology companies themselves. There are similar regulatory measures relevant to our analysis in Part 1, which include the broader GIFCT organisation and Tech Against Terrorism (the partnership between tech companies and the UN Counter-Terrorism Executive Directorate).

Increasingly, the Christchurch Call is likely to be invoked by non-state actors seeking to resist regulation by nation states and this is likely to test the long term durability of the call. Further, nation state signatories will be called upon to condemn actions by other nation states, to the extent these are inconsistent with the principles of the Call. When the Christchurch Call was negotiated, a group of civil society organisations raised a list of concerns about it.[95] This list of concerns is a useful resource for the investors in seeking to understand the relevant issues.

In this regard, the response protocols we identified in Part 1 are again another form of non-law regulatory response founded on partnerships between institutions. The Content Incident Protocol is also an example of a coordinated regulatory response that makes use of algorithmic tools (the shared hash database) as well as partnership and communication protocols among the platforms, and with nation states. Broader content moderation efforts made by the platforms as well as changes to their policies and internal processes can also fairly be described as regulatory responses, even if they are not reliant on the legal authority of nation states.

OVERSIGHT BOARD (FACEBOOK)

The UN Special Rapporteur for freedom of expression, having comprehensively reviewed the issues created by state regulation of social media platform, has called for an entity that would develop case law and act as a kind of "social media council" for elucidating how platforms are applying their content moderation policies:[96]

> *The companies are implementing "platform law", taking actions on content issues without significant disclosure about those actions. Ideally, companies should develop a kind of case law that would enable users, civil society and States to understand how the companies interpret and implement their standards. While such a "case law" system would not involve the kind of reporting the public expects from courts and administrative bodies, a detailed repository of cases and examples would clarify the rules much as case reporting does. A*

[95] See Civil Society Positions on Christchurch Call Pledge, available at <https://www.eff.org/files/2019/05/16/community_input_on_christchurch_call.pdf>.

[96] Above, A/HRC/38/35 at para 63.

> *social media council empowered to evaluate complaints across the ICT sector could be a credible and independent mechanism to develop such transparency.*

In 2019, Facebook announced it would establish an "Oversight Board", which we say closely resembles the kind of social media council endorsed by the Special Rapporteur. We have seen little recognition in public discussions of the way that the Oversight Board can be traced to human rights conclusions by a UN Special Rapporteur and we think this should be given greater weight when it comes to assessing the suitability of the Oversight Board as a regulatory solution.

The relevant features of the Oversight Board are as follows.[97]

- The Board's operations are framed by a founding charter, dated September 2019.
- The Board's founding charter requires the Board to consider its previous decisions as being of binding effect, in a similar way to a Court. The Board is required to explain its decisions in writing.
- The Board is also tasked with making policy recommendations, in a way that a strictly legalistic Court would not. Facebook can seek these directly from the Board, or the Board can make such recommendations in the course of its decisions.
- The Board can seek information from Facebook to assist the Board to make decisions and make recommendations, although to date Facebook has been resistant to complying with many of these requests for information.
- The Boards' members must number no less than 11 and are anticipated to reach around 40. The members enjoy a 3-year term and current members include people with highly prestigious qualifications and experience across human rights, law, policy, journalism, and academia.
- The intent of the Board is that it should refrain from implementing domestic laws by nation states when making its decisions. This meant, for example, in relation to the Board's decision about Facebook's suspension of President Trump, that the US Constitution was not directly relevant. Observers have speculated that this may amplify the strength of countries' reactions when Facebook acts against domestic authoritarian or extremist leaders within States.
- The charter requires the Board to have regard to the following sources of guidance, principle or law in making its decisions and recommendations:
    - Facebook's values
    - Facebook's content policies
    - Any prior board decisions "when the facts, applicable policies, or other factors are substantially similar"
    - Human rights norms protecting free expression
- The Oversight Board is funded by an endowment made by Facebook, so its funding is not contingent on Facebook's continued approval.
- The broad plan is that the Oversight Board could perform the same function that it does for Facebook for other platforms too.

[97] Key points drawn from the Board's founding charter at <https://oversightboard.com/governance/>.

The Oversight Board's charter states that the Board's resolutions of each case:[98]

> *… will be binding and Facebook will implement it promptly, unless implementation of a resolution could violate the law. In instances where Facebook identifies that identical content with parallel context — which the board has already decided upon — remains on Facebook, it will take action by analyzing whether it is technically and operationally feasible to apply the board's decision to that content as well. When a decision includes policy guidance or a policy advisory opinion, Facebook will take further action by analyzing the operational procedures required to implement the guidance, considering it in the formal policy development process of Facebook, and transparently communicating about actions taken as a result.*

In a number of the Board's decisions, it has turned to human rights doctrine and United Nations-related instruments in order to provide guidance for how it should make its own decisions. In the Trump decision, for example, it applied the Rabat Plan of Action's six-part threshold test for assessing whether freedom of expression should be restricted on the basis of incitement to hatred or violence.[99]

Some have framed the Board as a relatively self-interested attempt at self-regulation by Facebook in order to stave off regulation by states, but to reduce it to this purpose would be a mistake. That is because it would fail to account for the way that the Oversight Board is in many ways a suitable or even desirable broad solution to the issue of content moderation regulation. It also overlooks the fact that state-level regulation is complex, should not be rushed, and does not yet exist in many cases. Non-state regulation like the Oversight Board has the benefit of moving much faster and allowing for flexibility in operations as the Board's practice evolves.

There is a complex and important body of commentary on why the Oversight Board is limited as a regulatory response, but we think many of these criticisms can be dealt with over time as the practice and procedure of the Board develops. In any event, the eminence of the people appointed to membership of the Board, the long duration of the available funding, and the comparative absence of any alternative body which applies human rights instruments to offer both legal and policy guidance to platform companies, means that the Oversight Board will endure for the foreseeable future.

There is ample indication that the Oversight Board plans to use its position to procure more detailed information from Facebook about how its products operate, consistent with our overall conclusions that good regulation would start by inducing greater transparency and reporting. The Oversight Board has already asked Facebook to answer some tough questions, and it recommended "an open reflection on the design and policy choices … that may enable its platform to be abused" in relation to the events of 6 January 2021 at the US Capitol.

In a similar vein, the Board has recommended that Facebook make its data open to investigators and accountability mechanisms for any situation where "grave violations of international criminal, human rights and humanitarian law" are being investigated or prosecuted, consistent with our overall recommendation that transparency-oriented regulation

[98] Oversight Board charter, art 4.

[99] It is apparent that Facebook contributed to the formulation of Rabat plan, which has been cited in a range of international human rights instruments. See Trump Oversight Board decision 2021-001-FB-FBR at p 30.

is best practice at this stage in the platform regulation trajectory:[100]

> *Facebook has a responsibility to collect, preserve and, where appropriate, share information to assist in the investigation and potential prosecution of grave violations of international criminal, human rights and humanitarian law by competent authorities and accountability mechanisms. Facebook's corporate human rights policy should make clear the protocols the company has in place in this regard. The policy should also make clear how information previously public on the platform can be made available to researchers conducting investigations that conform with international standards and applicable data protection law.*

In its decision on Facebook's decision to suspend President Trump, the Oversight Board sought a range of information from Facebook to help inform its decision, which indicates the kind of information that transparency requirements might impose on content moderation platforms:[101]

> *In this case, the Board asked Facebook 46 questions, and Facebook declined to answer seven entirely, and two partially. The questions that Facebook did not answer included questions about how Facebook's news feed and other features impacted the visibility of Mr Trump's content; whether Facebook has researched, or plans to research, those design decisions in relation to the events of January 6, 2021; and information about violating content from followers of Mr. Trump's accounts. The Board also asked questions related to the suspension of other political figures and removal of other content; whether Facebook had been contacted by political officeholders or their staff about the suspension of Mr. Trump's accounts; and whether account suspension or deletion impacts the ability of advertisers to target the accounts of followers. Facebook stated that this information was not reasonably required for decision-making in accordance with the intent of the Charter; was not technically feasible to provide; was covered by attorney/client privilege; and/or could not or should not be provided because of legal, privacy, safety, or data protection concerns.*

In particular, the Board raised questions about the platform's design and engineering in response to Facebook's allegations that Trump had abused its platforms to create a narrative of election interference:

> *Facebook stated to the Board that it considered Mr. Trump's "repeated use of Facebook and other platforms to undermine confidence in the integrity of the election (necessitating repeated application by Facebook of authoritative labels correcting the misinformation) represented an extraordinary abuse of the platform." The Board sought clarification from Facebook about the extent to which the platform's design decisions, including algorithms, policies, procedures and technical features, amplified Mr Trump's posts after the election and whether Facebook had conducted any internal analysis of whether such design decisions may have contributed to the events of January 6. Facebook declined to answer these questions. This makes it difficult for the Board to assess whether less severe measures, taken earlier, may have been sufficient to protect the rights of others.*

This suggests that, while the Oversight Board proposal as a whole may be desirable, there is room for regulatory support for any situation where Facebook is being asked to disclose information, but it declines to do so.

[100] Case decision 2021-001-FB-FBR at p 37.

[101] Case decision 2021-001-FB-FBR at p 21.

Despite its willingness to ask difficult questions of Facebook, the Oversight Board has also explicitly directed Facebook to "resist pressure" from governments to silence political opposition.[102]

> *Restrictions on speech are often imposed by or at the behest of powerful state actors against dissenting voices and members of political oppositions. Facebook must resist pressure from governments to silence their political opposition. When assessing potential risks, Facebook should be particularly careful to consider the relevant political context. In evaluating political speech from highly influential users, Facebook should rapidly escalate the content moderation process to specialized staff who are familiar with the linguistic and political context and insulated from political and economic interference and undue influence. This analysis should examine the conduct of highly influential users off the Facebook and Instagram platforms to adequately assess the full relevant context of potentially harmful speech. Further, Facebook should ensure that it dedicates adequate resourcing and expertise to assess risks of harm from influential accounts globally.*

[102] Case decision 2021-001-FB-FBR at p 36.

CONCLUDING COMMENTS ON REGULATION

As a whole, we are persuaded by the weight of expert criticism that the likely effect of the current regulatory trajectory is highly concerning from a human rights perspective. Much of the proposed legislation creates compliance requirements which demonstrate that regulators believe content moderation is a simple rather than complex exercise; that accurate moderation at scale is possible within exceedingly short timeframes; and that automation can accurately and safely achieve this. None of this is correct.

There is a risk that "human rights" can be framed by states and by anti-platform advocates as being idealistic or inconvenient barriers to be avoided or overcome. By contrast, they must be seen as essential protective limitations for human dignity and flourishing. Human rights instruments and human rights law contain within them acceptable and proportionate ways to limit and balance human rights, but limiting human rights in this way should not be confused as avoiding or abandoning an overall commitment to a human rights approach.

Content-specific regulation approaches raise a host of practical and legal issues. It will take time to assess what effects any of this content-specific regulation will have. One feature of any legislation that imposes content-specific obligations is that it requires significant amounts of discretion, judgement, and the balancing of competing factors. That means its merits can only be assessed after it has been operational for a reasonable period of time.

Given the risk to individual human rights created by partnerships between platforms' digital infrastructure and the state's power, one insight we have taken from our research is that there are principled merits to non-state approaches to the regulation of expression. The primary benefit of non-state regulatory measures is that they leave final authority for determining what is permissible and impermissible expression to bodies other than the state. In this regard, the Oversight Board, established by Facebook, is a regulatory intervention for dealing with content moderation issues that deserves a fair chance.

States should limit themselves to intervening only in content which is already illegal, such as incitement to violence, threats of violence, or discrimination. States should resist the temptation to attempt to create a completely "safe" online environment without carefully defining what they mean by "safety", and considering the necessity, proportionality, and legality of their interventions to restrict freedom of expression in the name of the safety of others. All such interventions require clear and demonstrable justification using interventions which are capable of being reasonably understood and responded to by users, platforms, and governments.

It is important to take a long-term view when it comes to assessing how content moderation is regulated. One crucial dynamic to bear in mind is the way that freedom of expression is a right that protects individuals' ability to draw public attention to breaches of other human rights, including fundamentals like the right to life, the right to participate in public life and vote, the right to practice religious beliefs, the right against arbitrary detention, and the right to refuse medical treatment. The platforms add a novel dynamic to the interaction between States and individuals when it comes to freedom of expression. That is because they are digital infrastructures for moderating content that create bottlenecks for human interaction the likes of which human society has never seen. In this regard, it is essential to consider the way that, once legal powers to control these platforms are handed to states, the mechanisms for raising awareness about how that technology is abused are also suppressed. This could have a cascading effect over time which must be anticipated.

We think it is unrealistic to suggest that no form of algorithmic monitoring will be used to moderate content, particularly given the scale of the platforms. As such, regulation should reasonably anticipate the use of algorithmic monitoring systems. However, regulation should not assume that these systems will be accurate without human oversight. It should also not assume that they are always reliable. Regulation should anticipate the fact that imposing harsh penalties on platforms for failing to remove content rapidly will lead to heavier reliance on algorithmic systems, and the weighting of these systems toward over-removal (false positive) rather than under-removal (false negative) of content.

APPENDIX: ABOUT BRAINBOX AND ITS ROLE IN THIS PROJECT

Brainbox is an independent consultancy and think tank based in New Zealand, which specialises in issues at the intersection of technology, politics, law and policy. Brainbox and its key personnel have prepared funded legal research reports and advice on the following subjects:

- The implementation of the law in digital systems and the representation and implementation of legal instruments in machine executable languages (Legislation as Code);
- The use of algorithmic methods to analyse written decisions by judicial bodies and the policy implications of this, including methods of enhancing access to primary legal materials (Judgments as Data);
- The relationship between concepts of “trust” and “automated decision-making”, to support a wider research programme by the Digital Council for Aotearoa (Trust and Automated Decision Making);
- The legal implications of emerging technologies that create highly convincing but unreliable audio-visual media and how the New Zealand legal system deals with potentially harmful audio-visual content (Deepfakes and synthetic media);
- The policy implications of misinformation and disinformation, including attempts to regulate the creation and distribution of such information;
- A range of research investigations into health and disability policy, including how human rights instruments do or do not influence such policy (reports on accessibility, access to justice and human rights).

Brainbox’s brief was to apply its expertise in this and related subjects in order to reach conclusions and provide key insights to the group on the two key questions identified in parts 1 and 2. Brainbox’s role has not been to provide advice or recommendations to the group, whose members each have their own priorities and obligations when it comes to responsible investment practices and their relationships with the platform companies.

In this report, we limit our comments to the questions set out in the brief and aim to share insights to support the investor group to make its own decisions and recommendations. This includes any advocacy activity the group wishes to engage in with the platform companies or with state-level regulators. This report gives the investor group a basis from which to assess the likely implications and trade-offs of any recommendations they wish to make.

APPENDIX: EXPLAINING SCOPE OF ASSESSMENT IN PART 1

EXCLUDING QUESTIONS OF WHAT CAUSED THE REAL WORLD VIOLENCE

It is possible that the products offered by the platforms were a cause of the physical violence on 15 March to the extent that they were used to radicalised the individual and persuaded or inspired him to commit a violent act. They may also be contributing to real world violence elsewhere. We note that:

- The shooter self-reported that content he consumed on YouTube had an inspirational effect. The Report of the Royal Commission does not rule out that content accessed via YouTube may have had a radicalising effect on the Christchurch terrorist.
- There are empirical studies that examine the way that YouTube's recommendation algorithms recommend terrorist or extremist content. One study concluded that in 2016 the recommendation algorithm may have recommended such content, although by 2020 this situation was improved.[103]
- The genesis of the shared hash database now managed by GIFCT came after a series of terror attacks in 2015-2016 which were thought to have been inspired by online propaganda from ISIS intended to inspire lone actor attacks. This led to the EU Code of Conduct on Illegal Hate Speech and the auditing mechanisms under that instrument, and then to the founding of GIFCT in early 2017.
- It is important to note that, while YouTube is targeted most often when it comes to allegations of radicalisation, the Royal Commission also describes behaviour by the shooter on Facebook that might reasonably be argued to have contributed to his radicalisation to extremist views and to violence, and there is an extensive history of scholarship on the role of Twitter in spreading ISIS propaganda, which has as one of its goals the recruitment of future lone actor terrorists.

We initially attempted to separate what happened in the 15 March terror attacks into broadly 'online' and 'offline' components. The offline components would include the real-world violence and the online components would include the objectionable audio-visual content, as well as other aspects of the event that were internet-based. This would allow us to separately examine the platforms' contributions to both the online and offline components of the attack.

We concluded that the division between these two categories at times was difficult to maintain. In particular, there is a plausible relationship between the dissemination of online material produced during an attack and the consequent inspiration of future physical attacks, which then lead to further online content.[104]

[103] Murthy D, 'Evaluating Platform Accountability: Terrorist Content on YouTube' [2021] American Behavioral Scientist 0002764221989774.

[104] This can be seen in the shooter's allusions to Anders Breivik, as well as in subsequent attackers' allusions to the events of 15 March 2019.

However, we have excluded the question of the platforms' contribution to the real world violence on 15 March (as distinct from the online components of the attack). There are several reasons for this:

- Practically speaking, the contribution of social media platforms to radicalisation and radicalisation to violence is a growing body of empirical research which is beyond the scope of this project.
- Radicalisation is not the same as radicalisation to violence. While many people access radical content (including some videos hosted on YouTube), and some of these people develop radical beliefs, a comparatively smaller number go on to commit violent acts like the March 15 attacks. The Report of the Royal Commission especially noted this, based on its consultation with a range of experts and laypersons.
- The Report of the Royal Commission documents a range of factors which likely radicalised the individual, and which may have motivated him to commit violence. As such, it is difficult to accurately assess the degree to which content consumed on the Platforms was more or less causative than these other factors – and subsequently, the degree to which the platforms may need to take action to prevent future violent events.

The providence supporting the claim that YouTube content was the primary online source of radicalisation is the terrorist himself, primarily by way of a claim made in the manifesto. There is good reason to view him as an unreliable source. For example, it is well documented that the manifesto is riddled with strategic lies, misdirection, and attempts to incite further violence. Moreover, the Royal Commission acknowledges that the terrorist knew how to conceal his online activity through VPNs, TOR browsing, and encryption, meaning that nobody knows the full range of the terrorist's internet activity. As a result, it is plausible that YouTube played a lesser role in his radicalisation than he reports, and less so than other online and offline environments. It is plausible that the terrorist has an ulterior motive in singling out YouTube in his radicalisation. For example, he may be misdirecting attention away from 4Chan and 8Chan, which regularly host content and communications that is significantly more radical than what is permitted on YouTube.

EXCLUDING THE SHOOTER'S MANIFESTO

Overall, the brief for this report is directed toward objectionable audio-visual content, rather than the terrorist manifesto. This is not the only reason that we have excluded it from substantial analysis within this report:

- The manifesto is a text-based document. Even when in PDF (an image format), it is comparatively easy for automated content moderation systems to detect the presence of the manifesto. This may go some way to explaining why there seems to be less concern about its continued availability on the platforms.
- Though obviously distressing and classified as objectionable by New Zealand's Chief Censor, the manifesto is a materially different document than the audio-visual content. It is uncomfortable to describe it as less egregious than the audio-visual content, but we cannot avoid noting that it lacks some of the more objectionable elements. For example, it is not graphically violent, it does not contain personal information about the victims, and it is much less accessible than the audio-visual record of the attack.

- While there is no question that the manifesto is in breach of the voluntary standards of the platforms, it engages legal questions around balancing limitations on the right to convey and access information more than does the objectionable audio-visual content. For example, the manifesto is regularly studied outside of New Zealand, and is available within New Zealand by application to the Chief Censor.

In short, there are theoretical reasons why the presence of the manifesto on the platforms is a qualitatively different matter than the continued presence of the video content. These reasons influenced our shared decision with the investor group to exclude it from the scope of our assessment.

EXCLUDING OTHER PLATFORMS

The brief narrows focus to Google/YouTube, Facebook, and Twitter. A range of other internet platforms and websites had significant roles in reacting to the OCC which related to the 15 March attacks, but these fall outside the scope of our assessment. We note that a wide range of platforms are joining (or being encouraged to join) the international and inter-platform collaborative efforts we discuss in this report. The larger platforms are adopting a degree of responsibility for passing on best practice, and access to digital tools where appropriate, to aid content moderation practices for the less-resourced platforms. The investor group has also chosen to focus on these companies for strategic reasons which include:

- The investors' view that Facebook, Alphabet and Twitter are the key companies responsible for the main platforms where the Christchurch videos were distributed in a harmful way.
- They have the widest reach and the ability to make the investment required to mitigate spread.
- They are well placed to be influencers and first movers within the industry.

EXCLUDING NEW ZEALAND GOVERNMENT AGENCIES

The brief in Part 1 excludes matters relating to government agencies, consistent with the investors' roles as shareholding entities in the platforms. We note that the Report of the Royal Commission of Inquiry into the 15 March attacks is focussed almost entirely on the role and responsibility of government agencies. We have relied on that report for understanding the attacks. We do pause to note that the Royal Commission identified areas within government administration that could be improved and might prevent a future attack.

BIBLIOGRAPHY TO PART 1

'6.2m Tweets on EU Elections as Voters Turn to Twitter for Conversation' <https://blog.twitter.com/en_us/topics/company/2019/voters_turn_to_twitter_for_eu_elections.html> accessed 6 April 2021

'18 Trends That Highlight Fundamental Shifts in Culture' <https://blog.twitter.com/en_us/topics/insights/2019/18-trends-that-highlight-fundamental-shifts-in-culture.html> accessed 5 April 2021

'2019 El Paso Shooting', *Wikipedia* (2021) <https://en.wikipedia.org/w/index.php?title=2019_El_Paso_shooting&oldid=1015907913> accessed 6 April 2021

'A Further Update on New Zealand Terrorist Attack' (*About Facebook*, 21 March 2019) <https://about.fb.com/news/2019/03/technical-update-on-new-zealand/> accessed 15 March 2021

'A Look at the Research Behind Twitter Engage' <https://blog.twitter.com/en_us/a/2016/a-look-at-the-research-behind-twitter-engage.html> accessed 6 April 2021

'A Safer Internet for Europe, the Middle East and Africa' (*Google*, 11 February 2020) <https://blog.google/technology/safety-security/safer-internet-europe-middle-east-and-africa/> accessed 1 April 2021

'A Timeline of Recent Terrorist Attacks in Europe' (*Time*) <https://time.com/4607481/europe-terrorism-timeline-berlin-paris-nice-brussels/> accessed 1 April 2021

Abilov A and others, 'VoterFraud2020: A Multi-Modal Dataset of Election Fraud Claims on Twitter' [2021] arXiv:2101.08210 [cs] <http://arxiv.org/abs/2101.08210> accessed 6 April 2021

'About Tech Against Terrorism - Tech Against Terrorism' (4 September 2017) <https://www.techagainstterrorism.org/about/, https://www.techagainstterrorism.org/about/> accessed 18 March 2021

'Addressing Creator Feedback and an Update on My 2019 Priorities' (*blog.youtube*) <https://blog.youtube/inside-youtube/addressing-creator-feedback-and-update/> accessed 5 April 2021

'Addressing the Abuse of Tech to Spread Terrorist and Extremist Content' <https://blog.twitter.com/en_us/topics/company/2019/addressing-the-abuse-of-tech-to-spread-terrorist-and-extremist-c.html> accessed 6 April 2021

Alparslan Y and others, 'Towards Evaluating Gaussian Blurring in Perceptual Hashing as a Facial Image Filter' [2020] arXiv:2002.00140 [cs] <http://arxiv.org/abs/2002.00140> accessed 6 April 2021

Amarasingam DA, 'Turning the Tap Off: The Impacts of Social Media Shutdown After Sri Lanka's Easter Attacks' (*GNET*) <https://gnet-research.org/2021/03/05/turning-the-tap-off-the-impacts-of-social-media-shutdown-after-sri-lankas-easter-attacks/> accessed 13 April 2021

'An Update on Combating Hate and Dangerous Organizations' (*About Facebook*, 12 May 2020) <https://about.fb.com/news/2020/05/combating-hate-and-dangerous-organizations/> accessed 9 March 2021

‘An Update on Our Commitment to Fight Terror Content Online’ (*blog.youtube*) <https://blog.youtube/news-and-events/an-update-on-our-commitment-to-fight-terror/> accessed 31 March 2021

‘An Update on Our Efforts to Combat Terrorism Online’ (*About Facebook*, 20 December 2019) <https://about.fb.com/news/2019/12/counterterrorism-efforts-update/> accessed 25 March 2021

‘An Update on Our Efforts to Combat Violent Extremism’ <https://blog.twitter.com/en_us/a/2016/an-update-on-our-efforts-to-combat-violent-extremism.html> accessed 6 April 2021

‘Announcing Google.Org’s New Safety Grants in Europe’ (*Google*, 4 February 2020) <https://blog.google/outreach-initiatives/google-org/announcing-googleorgs-new-safety-grants-europe/> accessed 1 April 2021

‘Arrested Coast Guard Officer Allegedly Planned Attack “On A Scale Rarely Seen”’ (*NPR.org*) <https://www.npr.org/2019/02/20/696470366/arrested-coast-guard-officer-planned-mass-terrorist-attack-on-a-scale-rarely-see> accessed 11 March 2021

Awan I, ‘Cyber-Extremism: Isis and the Power of Social Media’ (2017) 54 Society 138

Azarafrooz A and Brock J, ‘Fuzzy Hashing as Perturbation-Consistent Adversarial Kernel Embedding’ [2018] arXiv:1812.07071 [cs, stat] <http://arxiv.org/abs/1812.07071> accessed 6 April 2021

Baldi M and others, ‘On Fuzzy Syndrome Hashing with LDPC Coding’ [2011] arXiv:1107.1600 [cs, math] <http://arxiv.org/abs/1107.1600> accessed 6 April 2021

Banchik AV, ‘Disappearing Acts: Content Moderation and Emergent Practices to Preserve at-Risk Human Rights–Related Content’ [2020] New Media & Society 1461444820912724

Basra R, ‘The YouTube Browsing Habits of a Lone-Actor Terrorist’ (*GNET*) <https://gnet-research.org/2020/06/22/the-youtube-browsing-habits-of-a-lone-actor-terrorist/> accessed 13 April 2021

‘Bear Witness, Take Action’ (*blog.youtube*) <https://blog.youtube/news-and-events/bear-witness-take-action/> accessed 1 April 2021

Benesch S, ‘But Facebook’s Not a Country: How to Interpret Human Rights Law for Social Media Companies’ (2020) 38 Yale Journal on Regulation Bulletin 86

Bhaskara VS and Bhattacharyya D, ‘Emulating Malware Authors for Proactive Protection Using GANs over a Distributed Image Visualization of Dynamic File Behavior’ [2018] arXiv:1807.07525 [cs, stat] <http://arxiv.org/abs/1807.07525> accessed 6 April 2021

‘Big Tent Sendai: Smarter Ways to Share Information in a Crisis’ (*Google*, 3 July 2012) <https://blog.google/outreach-initiatives/google-org/big-tent-sendai-smarter-ways-to-share/> accessed 1 April 2021

Bishop DP, ‘Online Terrorist Content: Is It Time for an Independent Regulator?’ (*GNET*) <https://gnet-research.org/2020/11/16/online-terrorist-content-is-it-time-for-an-independent-regulator/> accessed 13 April 2021

Biswas R and others, ‘Perceptual Hashing Applied to Tor Domains Recognition’ [2020] arXiv:2005.10090 [cs] <http://arxiv.org/abs/2005.10090> accessed 6 April 2021

Blake S, 'Embedded Blockchains: A Synthesis of Blockchains, Spread Spectrum Watermarking, Perceptual Hashing & Digital Signatures' [2020] arXiv:2009.00951 [cs, math] <http://arxiv.org/abs/2009.00951> accessed 6 April 2021

'Bot or Not? The Facts about Platform Manipulation on Twitter' <https://blog.twitter.com/en_us/topics/company/2020/bot-or-not.html> accessed 6 April 2021

Brad Smith, 'Microsoft, Other Tech Industry Leaders Team up with an International Coalition of Governments for a Multi-Stakeholder Solution' (*Microsoft On the Issues*, 24 September 2019) <https://blogs.microsoft.com/on-the-issues/2019/09/23/microsoft-other-tech-industry-leaders-team-up-with-an-international-coalition-of-governments-for-a-multi-stakeholder-solution/> accessed 8 March 2021

'Bringing Facebook Live to Android and More Countries' (*About Facebook*, 26 February 2016) <https://about.fb.com/news/2016/02/bringing-facebook-live-to-android-and-more-countries/> accessed 24 March 2021

'Bringing New Redirect Method Features to YouTube' (*blog.youtube*) <https://blog.youtube/news-and-events/bringing-new-redirect-method-features/> accessed 31 March 2021

'Bringing the Campaigning Power of Twitter and Counter-Narratives to Spain' <https://blog.twitter.com/en_us/a/2016/bringing-the-campaigning-power-of-twitter-and-counter-narratives-to-spain.html> accessed 5 April 2021

'Building a Safer Internet - Google Safety Center' <https://safety.google/engineering-center/> accessed 1 April 2021

'Building a Safer Internet, from Europe to Africa' (*Google*, 9 February 2021) <https://blog.google/technology/safety-security/building-safer-internet-europe-africa/> accessed 1 April 2021

'Christchurch Earthquake — One Year Later: Live Streaming the Memorial Service on YouTube' (*blog.youtube*) <https://blog.youtube/news-and-events/christchurch-earthquake-one-year-later/> accessed 31 March 2021

'Christchurch Mosque Attack Livestream : Featured Classification Decisions : OFLC' <https://www.classificationoffice.govt.nz/news/featured-classification-decisions/christchurch-mosque-attack-livestream/> accessed 18 March 2021

'Civil Society Positions on Christchurch Call Pledge' <https://www.eff.org/files/2019/05/16/community_input_on_christchurch_call.pdf> accessed 6 April 2021

'Code of Practice on Disinformation | Shaping Europe's Digital Future' <https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation> accessed 1 April 2021

Coll S, 'Alex Jones, the First Amendment, and the Digital Public Square' (*The New Yorker*) <https://www.newyorker.com/magazine/2018/08/20/alex-jones-the-first-amendment-and-the-digital-public-square> accessed 31 March 2021

'Combating Hate and Extremism' (*About Facebook*, 17 September 2019) <https://about.fb.com/news/2019/09/combating-hate-and-extremism/> accessed 25 March 2021

'Combating Violent Extremism' <https://blog.twitter.com/en_us/a/2016/combating-violent-extremism.html> accessed 6 April 2021

Comerford M, 'Two Years On: Understanding the Resonance of the Christchurch Attack on Imageboard Sites' (*GNET*) <https://gnet-research.org/2021/03/24/two-years-on-understanding-the-resonance-of-the-christchurch-attack-on-imageboard-sites/> accessed 8 April 2021

‘Community Standards’ <https://m.facebook.com/communitystandards/additional_information/> accessed 1 April 2021

Content Standards Forum (Facebook, 26 March 2019) <https://about.fb.com/wp-content/uploads/2018/11/csf-final-deck_03.26.19.pdf> accessed 1 April 2021

Conway M and others, ‘Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts’ (2019) 42 Studies in Conflict & Terrorism 141

Covington P, Adams J and Sargin E, ‘Deep Neural Networks for YouTube Recommendations’, *Proceedings of the 10th ACM Conference on Recommender Systems* (ACM 2016) <https://dl.acm.org/doi/10.1145/2959100.2959190> accessed 6 April 2021

‘Creating a Dataset and a Challenge for Deepfakes’ <https://ai.facebook.com/blog/deepfake-detection-challenge/> accessed 25 March 2021

‘Crisis Response’ (*GIFCT*) <https://gifct.org/crisis-communications/> accessed 5 April 2021

Dalins J, Wilson C and Boudry D, ‘PDQ & TMK + PDQF -- A Test Drive of Facebook’s Perceptual Hashing Algorithms’ [2019] arXiv:1912.07745 [cs] <http://arxiv.org/abs/1912.07745> accessed 6 April 2021

D’Anastasio C, ‘A Christchurch Report Points to YouTube’s Radicalization Trap’ *Wired* <https://www.wired.com/story/christchurch-shooter-youtube-radicalization-extremism/> accessed 13 April 2021

Davidson J and others, ‘The YouTube Video Recommendation System’ (2010)De Herve JDG and others, ‘A Perceptual Hash Function to Store and Retrieve Large Scale DNA Sequences’ [2014] arXiv:1412.5517 [cs, q-bio] <http://arxiv.org/abs/1412.5517> accessed 6 April 2021

‘December 2020 Coordinated Inauthentic Behavior Report’ (*About Facebook*, 12 January 2021) <https://about.fb.com/news/2021/01/december-2020-coordinated-inauthentic-behavior-report/> accessed 25 March 2021

‘Defending the Truth of the Holocaust in 2021’ (*blog.youtube*) <https://blog.youtube/news-and-events/defending-the-truth-holocaust-2021/> accessed 31 March 2021

‘Designing with Constraint: Twitter’s Approach to Email’ <https://blog.twitter.com/en_us/a/2015/designing-with-constraint-twitters-approach-to-email.html> accessed 5 April 2021

‘Despite A Ban, Facebook Continued To Label People As Interested In Militias For Advertisers’ (*BuzzFeed News*) <https://www.buzzfeednews.com/article/ryanmac/facebook-militia-interest-category-advertisers-ban> accessed 9 April 2021

‘Dialogue with Sen. Lieberman on Terrorism Videos’ (*blog.youtube*) <https://blog.youtube/news-and-events/dialogue-with-sen-lieberman-on/> accessed 31 March 2021

‘Discord Chats May Be Crucial to Lawsuits over Neo-Nazi Violence’ (*Engadget*) <https://www.engadget.com/2017-08-26-discord-chats-may-help-charlottesville-lawsuits.html> accessed 11 March 2021

Douek E, ‘Facebook’s “Oversight Board:” Move Fast With Stable Infrastructure And Humility’ (2019) 21

——, ‘Australia’s “Abhorrent Violent Material” Law: Shouting “Nerd Harder” and Drowning Out Speech’ (Social Science Research Network 2019) SSRN Scholarly Paper ID 3443220 <https://papers.ssrn.com/abstract=3443220> accessed 6 April 2021

——, 'The Free Speech Blind Spot: Foreign Election Interference On Social Media' [2020] Draft – Combating Election Interference When Foreign Powers Target Democracies (Duncan B. Hollis & Jens David Ohlin eds., Oxford University Press, forthcoming 2020) 27

——, 'The Limits of International Law in Content Moderation' [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3709566> accessed 31 March 2021

——, 'The Rise of Content Cartels' [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3572309> accessed 31 March 2021

——, 'Governing Online Speech: From "Posts-As-Trumps" To Proportionality & Probability'

Echikson W and Knodt O, 'Germany's NetzDG: A Key Test for Combatting Online Hate' (Social Science Research Network 2018) SSRN Scholarly Paper ID 3300636 <https://papers.ssrn.com/abstract=3300636> accessed 6 April 2021

Elhai W, 'Regulating Digital Harm Across Borders: Exploring a Content Platform Commission', *International Conference on Social Media and Society* (Association for Computing Machinery 2020) <https://doi.org/10.1145/3400806.3400832> accessed 5 April 2021

'Empowering Dutch NGOs to Amplify Their Voice on Twitter' <https://blog.twitter.com/en_us/a/2016/empowering-dutch-ngos-to-amplify-their-voice-on-twitter.html> accessed 5 April 2021

Erin Saltman, 'Countering Terrorism and Violent Extremism at Facebook: Technology, Expertise and Partnerships' in Maya Mirchandani (ed), *Tackling Insurgent Ideologies in a Pandemic World* (2020)

'European Council Conclusions on Security and Defence, 22/06/2017' <https://www.consilium.europa.eu/en/press/press-releases/2017/06/22/euco-security-defence/> accessed 1 April 2021

'Evaluating Platform Accountability: Terrorist Content on YouTube - Dhiraj Murthy, 2021' <https://journals-sagepub-com.ezproxy.otago.ac.nz/doi/10.1177/0002764221989774> accessed 6 April 2021

'Explainers' (*GNET*) <https://gnet-research.org/explainers/> accessed 18 March 2021

'Exposed Email Logs Show 8kun Owner in Contact With QAnon Influencers and Enthusiasts' (*bellingcat*, 7 January 2021) <https://www.bellingcat.com/news/2021/01/07/exposed-email-logs-show-8kun-owner-in-contact-with-qanon-influencers-and-enthusiasts/> accessed 11 March 2021

Facebook and others, 'One Dead, Three Injured in Poway Synagogue Shooting' (*San Diego Union-Tribune*, 27 April 2019) <https://www.sandiegouniontribune.com/news/public-safety/story/2019-04-27/reports-of-several-people-shot-at-poway-synagogue> accessed 11 March 2021

'Facebook at UNGA 2020' (*About Facebook*, 21 September 2020) <https://about.fb.com/news/2020/09/facebook-at-unga-2020/> accessed 25 March 2021

'Facebook Joins Other Tech Companies to Support the Christchurch Call to Action' (*About Facebook*, 15 May 2019) <https://about.fb.com/news/2019/05/christchurch-call-to-action/> accessed 9 March 2021

'Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism' (*About Facebook*, 26 June 2017) <https://about.fb.com/news/2017/06/global-internet-forum-to-counter-terrorism/> accessed 1 April 2021

'——' (*blog.youtube*) <https://blog.youtube/news-and-events/facebook-microsoft-twitter-and-youtube/> accessed 1 April 2021

'Facebook Says New Rule Would Have Stopped Christchurch Shooter Livestreaming' (*Stuff*, 15 May 2019) <https://www.stuff.co.nz/national/politics/112756865/facebook-says-new-rule-would-have-stopped-christchurch-shooter-livestreaming> accessed 14 April 2021

'Facebook Says No One Flagged NZ Mosque Shooting Livestream' (*The Salt Lake Tribune*) <https://sltrib.com/news/nation-world/2019/03/19/facebook-says-no-one> accessed 18 March 2021

'Facebook's Community Standards: How and Where We Draw the Line' (*About Facebook*, 23 May 2017) <https://about.fb.com/news/2017/05/facebooks-community-standards-how-and-where-we-draw-the-line/> accessed 24 March 2021

'Five Tips for Brands from #Twitter4Politics' <https://blog.twitter.com/en_us/a/2015/five-tips-for-brands-from-twitter4politics.html> accessed 6 April 2021

Ford P, 'Combatting Terrorist Propaganda' (2020) 15 Journal of Policing, Intelligence and Counter Terrorism 175

'Four Steps We're Taking Today to Fight Terrorism Online' (*Google*, 18 June 2017) <https://blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/> accessed 31 March 2021

'Frequently Asked Questions (FAQ) - Tech Against Terrorism' (28 November 2017) <https://www.techagainstterrorism.org/about/faq/, https://www.techagainstterrorism.org/about/faq/> accessed 18 March 2021

'From Countering Radicalization to Disrupting Illicit Networks: What's next for Google Ideas' (*Official Google Blog*) <https://googleblog.blogspot.com/2012/04/from-countering-radicalization-to.html> accessed 1 April 2021

Ganesh B and Bright J, 'Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation' (2020) 12 Policy & Internet 6

——, Extreme Digital Speech: Contexts, Responses, and Solutions (VOX-Pol Network of Excellence 2020)

'Getting Input on an Oversight Board' (*About Facebook*, 1 April 2019) <https://about.fb.com/news/2019/04/input-on-an-oversight-board/> accessed 5 April 2021

'GIFCT Transparency Report, July 2020' <https://gifct.org/wp-content/uploads/2020/10/GIFCT-Transparency-Report-July-2020-Final.pdf> accessed 5 April 2021

'Global Internet Forum to Counter Terrorism: An Update on Our Progress' (*blog.youtube*) <https://blog.youtube/news-and-events/global-internet-forum-to-counter/> accessed 1 April 2021

'Global Internet Forum To Counter Terrorism: An Update on Our Progress Two Years On' (*About Facebook*, 25 July 2019) <https://about.fb.com/news/2019/07/global-internet-forum-to-counter-terrorism-an-update-on-our-progress-two-years-on/> accessed 25 March 2021

'Global Internet Forum to Counter Terrorism to Hold First Meeting' <https://blog.twitter.com/en_us/topics/insights/2017/Global-Internet-Forum-to-Counter-Terrorism-to-hold-first-meeting.html> accessed 6 April 2021

‘Global Internet Forum to Counter Terrorism to Hold First Meeting in San Francisco’ (*About Facebook*, 31 July 2017) <https://about.fb.com/news/2017/07/global-internet-forum-to-counter-terrorism-to-hold-first-meeting-in-san-francisco/> accessed 25 March 2021

‘——’ (*About Facebook*, 31 July 2017) <https://about.fb.com/news/2017/07/global-internet-forum-to-counter-terrorism-to-hold-first-meeting-in-san-francisco/> accessed 5 April 2021

‘Global Internet Forum to Counter Terrorism to Hold First Meeting in San Francisco’ (*blog.youtube*) <https://blog.youtube/news-and-events/global-internet-forum-san-francisco/> accessed 1 April 2021

‘Global Network Initiative Releases 2015 Assessment Report’ (*Google*, 8 July 2016) <https://blog.google/outreach-initiatives/public-policy/global-network-initiative-releases-2015/> accessed 1 April 2021

‘GNI Resignation Letter’ (*Electronic Frontier Foundation*, 9 October 2013) <https://www.eff.org/document/gni-resignation-letter> accessed 6 April 2021

‘Google Ideas: Joining the Fight against Drug Cartels and Other Illicit Networks’ (*Google*, 16 July 2012) <https://blog.google/alphabet/google-ideas-joining-fight-against-drug/> accessed 1 April 2021

‘Google Ideas Launches Summit Against Violent Extremism’ (*Google Europe Blog*) <https://europe.googleblog.com/2011/06/google-ideas-launches-summit-against.html> accessed 1 April 2021

‘Google-Funded Report on White Supremacy Downplays YouTube’s Role in Driving People to Extremism’ (*Stuff*, 16 December 2020) <https://www.stuff.co.nz/technology/300185901/googlefunded-report-on-white-supremacy-downplays-youtubes-role-in-driving-people-to-extremism> accessed 5 April 2021

‘Google.Org Impact Challenge on Safety’ (*Google.org Impact Challenge on Safety*) <https://impactchallenge.withgoogle.com/safety2019> accessed 1 April 2021

Gorwa R, ‘What Is Platform Governance?’ (2019) 22 Information, Communication & Society 854

——, ‘The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content’ (2019) 8 Internet Policy Review <https://policyreview.info/node/1407> accessed 6 April 2021

Gorwa R, Binns R and Katzenbach C, ‘Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance’ (2020) 7 Big Data & Society 2053951719897945

Greenfield S, ‘Social Media Platforms: Preserving Evidence of International Crimes Notes’ (2018) 2 International Comparative, Policy & Ethics Law Review 821

Grimmelmann J, ‘The Virtues of Moderation’ (LawArXiv 2017) preprint <https://osf.io/qwxf5> accessed 6 April 2021

‘GSEC Dublin: A Content Responsibility Center for Europe’ (*Google*, 27 January 2021) <https://blog.google/around-the-globe/google-europe/gsec-dublin-content-responsibility-center-europe/> accessed 1 April 2021

‘Guest Post: Why We Must Remember the Holocaust’ <https://blog.twitter.com/en_us/topics/events/2019/we_remember.html> accessed 6 April 2021

Hans Bredow Institute, 'Setting Rules for 2.7 Billion: A (First) Look into Facebook's Norm-Making System: Results of a Pilot Study' (January 2020) <https://www.hans-bredow-institut.de/uploads/media/Publikationen/cms/media/7mkl6yl_AP_WiP001InsideFacebook.pdf> accessed 1 April 2021

'Hard Questions: How We Counter Terrorism' (*About Facebook*, 15 June 2017) <https://about.fb.com/news/2017/06/how-we-counter-terrorism/> accessed 25 March 2021

Heidi Tworek, 'Social Media Councils' (*Centre for International Governance Innovation*, 28 October 2019) <https://www.cigionline.org/articles/social-media-councils> accessed 14 April 2021

Heller B, 'Combating Terrorist-Related Content Through AI and Information Sharing' (The Carr Center for Human Rights Policy, Harvard University 2019)

'How Twitter Is Fighting Spam and Malicious Automation' <https://blog.twitter.com/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html> accessed 6 April 2021

'How We're Supporting Smart Regulation and Policy Innovation in 2019' (*Google*, 8 January 2019) <https://blog.google/perspectives/kent-walker/principles-evolving-technology-policy-2019/> accessed 5 April 2021

'How YouTube Supports Elections' (*blog.youtube*) <https://blog.youtube/news-and-events/how-youtube-supports-elections/> accessed 5 April 2021

'Inclusion & Diversity Report May 2019' <https://blog.twitter.com/en_us/topics/company/2019/Board-Update-Inclusion-Diversity-Report-May2019.html> accessed 6 April 2021

'#Influencer Voices: How the National Association of Manufacturers Captures Attention on Twitter during Major Political Events' <https://blog.twitter.com/en_us/a/2015/influencer-voices-how-the-national-association-of-manufacturers-captures-attention-on-twitter.html> accessed 5 April 2021

'Insights from the 17th Twitter Transparency Report' <https://blog.twitter.com/en_us/topics/company/2020/ttr-17.html> accessed 6 April 2021

'Introducing Event Targeting' <https://blog.twitter.com/en_us/a/2015/introducing-event-targeting.html> accessed 5 April 2021

'Introducing Live Video and Collages' (*About Facebook*, 3 December 2015) <https://about.fb.com/news/2015/12/introducing-live-video-and-collages/> accessed 24 March 2021

'Introducing the New Twitter Transparency Center' <https://blog.twitter.com/en_us/topics/company/2020/new-transparency-center.html> accessed 6 April 2021

'Investigating Information Operations in West Papua: A Digital Forensic Case Study of Cross-Platform Network Analysis' (*bellingcat*, 11 October 2019) <https://www.bellingcat.com/news/rest-of-world/2019/10/11/investigating-information-operations-in-west-papua-a-digital-forensic-case-study-of-cross-platform-network-analysis/> accessed 1 April 2021

Jackson S, 'The Double-Edged Sword of Banning Extremists from Social Media' <https://osf.io/preprints/socarxiv/2g7yd/> accessed 6 April 2021

Jie Z, 'A Novel Block-DCT and PCA Based Image Perceptual Hashing Algorithm' [2013] arXiv:1306.4079 [cs] <http://arxiv.org/abs/1306.4079> accessed 6 April 2021

‘Jigsaw’ (*Jigsaw*) <https://jigsaw.google.com/> accessed 1 April 2021

‘Jimmy Wales on Systems and Incentives (Ep. 109)’ <https://conversationswithtyler.com/episodes/jimmy-wales/> accessed 24 March 2021

‘Joint Letter to New Executive Director, Global Internet Forum to Counter Terrorism’ (*Human Rights Watch*, 30 July 2020) <https://www.hrw.org/news/2020/07/30/joint-letter-new-executive-director-global-internet-forum-counter-terrorism> accessed 13 April 2021

@jshermcyber, ‘The Christchurch Report Points to Better Avenues for Internet Reform’ (*Lawfare*, 26 March 2021) <https://www.lawfareblog.com/christchurch-report-points-better-avenues-internet-reform> accessed 9 April 2021

Kate Klonick, ‘Facebook Released Its Content Moderation Rules. Now What?’

——, ‘A “Creepy” Assignment: Pay Attention to What Strangers Reveal in Public’

Kate Klonick and Thomas Kadri, ‘How to Make Facebook’s “Supreme Court” Work’

Katzenbach C and Ulbricht L, ‘Algorithmic Governance’ (2019) 8 Internet Policy Review <https://policyreview.info/node/1424> accessed 6 April 2021

Keen F, ‘Online Subcultures and the Challenges of Moderation’ (*GNET*) <https://gnet-research.org/2020/10/01/online-subcultures-and-the-challenges-of-moderation/> accessed 13 April 2021

‘Keeping Our Users Secure’ <https://blog.twitter.com/en_us/a/2013/keeping-our-users-secure.html> accessed 5 April 2021

Keller D, ‘Making Google the Censor’ (*New York Times (Online)*, 12 June 2017) <http://search.proquest.com/docview/1908292276/abstract/3629B3851B694C5FPQ/1> accessed 31 March 2021

——, ‘Don’t Export Restrictions On Speech’ *The New York times* (2018) A19

——, ‘Don’t Force Google to Export Other Countries’ Laws’ (*New York Times (Online)*, 10 September 2018) <http://search.proquest.com/docview/2101581443/abstract/787930A6ED0444B4PQ/1> accessed 31 March 2021

——, ‘The Stubborn, Misguided Myth That Internet Platforms Must Be “Neutral”’ *The Washington post* (Washington, DC, 2019)

——, ‘Facebook Filters, Fundamental Rights, and the CJEU’s Glawischnig-Piesczek Ruling’ (2020) 69 GRUR International 616

Keller D and Brown BD, ‘Europe’s Web Privacy Rules: Bad for Google, Bad for Everyone’ (*New York Times (Online)*, 25 April 2016) <http://search.proquest.com/docview/1783859103/abstract/F9541C172E5E42E0PQ/1> accessed 31 March 2021

Klonick K, ‘Re-Shaming the Debate: Social Norms, Shame and Regulation in an Internet Age’ (2016) 75 Maryland law review (1936) 1029

——, ‘Networked Technologies’ Transformation of Social Norms, Private Self-Regulation, and the Law’ (ProQuest Dissertations Publishing 2018) <https://search.proquest.com/docview/2088928846?pq-origsite=primo> accessed 6 April 2021

——, ‘The New Governors: The People, Rules, And Processes Governing Online Speech’ (2018) 131 Harvard law review 1598

——, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' [2020] the yale law journal 83

——, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2020) 129 The Yale law journal 2418

——, 'Content Moderation Modulation: Deliberating on How to Regulate--or Not Regulate--Online Speech in the Era of Evolving Social Media' (2021) 64 Communications of the ACM 29

——, 'Inside the Making of Facebook's Supreme Court' (*The New Yorker*) <https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court> accessed 31 March 2021

——, 'Inside the Team at Facebook That Dealt with the Christchurch Shooting' (*The New Yorker*) <https://www.newyorker.com/news/news-desk/inside-the-team-at-facebook-that-dealt-with-the-christchurch-shooting> accessed 31 March 2021

Klonick K and Kadri T, 'How to Make Facebook's "Supreme Court" Work' (*New York Times (Online)*, 17 November 2018) <http://search.proquest.com/docview/2134340716/abstract/6A651F6F91E04E97PQ/1> accessed 31 March 2021

'Korero Whakamauahara - Hate Speech (New Zealand Human Rights Commission, 2019)' <https://www.hrc.co.nz/files/2915/7653/6167/Korero_Whakamauahara-_Hate_Speech_FINAL_13.12.2019.pdf> accessed 8 March 2021

Langvardt K, 'Regulating Online Content Moderation' (2018) 106 The Georgetown Law Journal 1353

Leader Maynard J and Benesch S, 'Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention' (2016) 9 Genocide studies and prevention 70

Leman J and Pektaş Ş, *Militant Jihadism: Today and Tomorrow* (Leuven University Press 2019)

Lewis R, 'Broadcasting the Reactionary Right on YouTube' (Data & Society Research Institute)

Li K, 'Hashing for Multimedia Similarity Modeling and Large-Scale Retrieval' (University of Central Florida 2017)

Li Y, Jang J and Ou X, 'Topology-Aware Hashing for Effective Control Flow Graph Similarity Analysis' [2020] arXiv:2004.06563 [cs] <http://arxiv.org/abs/2004.06563> accessed 6 April 2021

Lyons M, 'Regulating Terrorist Content Online: Considerations and Trade-Offs' (*Counter Terror Business*, 14 October 2019) <https://counterterrorbusiness.com/features/regulating-terrorist-content-online-considerations-and-trade-offs> accessed 6 April 2021

'Machine Learning Can Identify Weapons in the Christchurch Attack Video' <https://www.vice.com/en/article/xwnzz4/machine-learning-artificial-intelligence-christchurch-attack-video-facebook-amazon-rekognition> accessed 18 March 2021

'Making It Easier to Report Threats to Law Enforcement' <https://blog.twitter.com/en_us/a/2015/making-it-easier-to-report-threats-to-law-enforcement.html> accessed 6 April 2021

'Making Our Rules Easier to Understand' <https://blog.twitter.com/en_us/topics/company/2019/rules-refresh.html> accessed 6 April 2021

'Mark Warner Is Ready to Fight for Section 230 Reform' (*Protocol — The people, power and politics of tech*, 22 March 2021) <https://www.protocol.com/policy/mark-warner-section-230> accessed 6 April 2021

Mark Zuckerberg, 'Some Thoughts on Facebook and the Election' (13 November 2016) <https://www.facebook.com/zuck/posts/10103253901916271> accessed 24 March 2021

'Mark Zuckerberg Says Fake News On Facebook Didn't Change The Election' (*BuzzFeed News*) <https://www.buzzfeednews.com/article/stephaniemlee/zuckerberg-techonomy-fake-news-election> accessed 24 March 2021

'Mass Violence, Extremism, and Digital Responsibility' (*U.S. Senate Committee on Commerce, Science, & Transportation*, 18 September 2019) <https://www.commerce.senate.gov/2019/9/mass-violence-extremism-and-digital-responsibility> accessed 6 April 2021

'Mastodon' <https://joinmastodon.org/> accessed 11 March 2021

Mattheis A, 'Beyond the "LULZ:" Memifying Murder as "Meaningful" Gamification in Far-Right Content' (*GNET*) <https://gnet-research.org/2021/01/18/beyond-the-lulz-memifying-murder-as-meaningful-gamification-in-far-right-content/> accessed 13 April 2021

Mayer J, 'Content Moderation for End-to-End Encrypted Messaging'

'Meet the Teams Keeping Our Corner of the Internet Safer' (*Google*, 5 February 2019) <https://blog.google/around-the-globe/google-europe/meet-teams-keeping-our-corner-internet-safer/> accessed 1 April 2021

'Microsoft Partners with Institute for Strategic Dialogue and NGOs to Discourage Online Radicalization to Violence' (*Microsoft On the Issues*, 18 April 2017) <https://blogs.microsoft.com/on-the-issues/2017/04/18/microsoft-partners-institute-strategic-dialogue-ngos-discourage-online-radicalization-violence/> accessed 1 April 2021

Montgomery M, 'Disinformation as a Wicked Problem: Why We Need Co-Regulatory Frameworks' 14

Murthy D, 'Evaluating Platform Accountability: Terrorist Content on YouTube' [2021] American Behavioral Scientist 0002764221989774

'/N/ - New Zealand Mobile Carriers Block 8chan, 4chan, and LiveLeak' (18 March 2019) <https://web.archive.org/web/20190318033153/https:/8ch.net/n/res/756614.html> accessed 11 March 2021

Nast C, 'Facebook's Supreme Court' (*The New Yorker*) <https://www.newyorker.com/podcast/political-scene/facebooks-supreme-court> accessed 31 March 2021

——, 'The Supreme Court of Facebook' (*The New Yorker*) <https://www.newyorker.com/podcast/the-new-yorker-radio-hour/the-supreme-court-of-facebook> accessed 31 March 2021

'New Progress in Using AI to Detect Harmful Content' <https://ai.facebook.com/blog/community-standards-report/> accessed 13 April 2021

'Next Steps for the Global Internet Forum to Counter Terrorism' (*About Facebook*, 23 September 2019) <https://about.fb.com/news/2019/09/next-steps-for-gifct/> accessed 25 March 2021

Nguyen DT and others, ‘Automatic Image Filtering on Social Networks Using Deep Learning and Perceptual Hashing During Crises’ [2017] arXiv:1704.02602 [cs] <http://arxiv.org/abs/1704.02602> accessed 6 April 2021

‘Operating with Impunity - Hateful Extremism: The Need for a Legal Framework’ (Commission for Countering Extremism 2021)

‘Opinion | How to Make Facebook’s “Supreme Court” Work - The New York Times’ <https://www.nytimes.com/2018/11/17/opinion/facebook-supreme-court-speech.html> accessed 6 April 2021

Osnos E, ‘How to Talk About the New Zealand Massacre: More Sunlight, Less Oxygen’ (*The New Yorker*) <https://www.newyorker.com/news/daily-comment/how-to-talk-about-the-new-zealand-massacre-more-sunlight-less-oxygen> accessed 31 March 2021

‘Our #DataForGood Partnership with New Zealand’s NCPACS’ <https://blog.twitter.com/en_us/topics/company/2020/christchurch-otago-nspacs.html> accessed 6 April 2021

‘Our Ongoing Work to Tackle Hate’ (*blog.youtube*) <https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/> accessed 9 March 2021

‘Oversight Frameworks for Content-Sharing Platforms’ (*Google*, 19 June 2019) <https://blog.google/outreach-initiatives/public-policy/oversight-frameworks-content-sharing-platforms/> accessed 5 April 2021

Pandey P, ‘One Year Since the Christchurch Call to Action: A Review’ (Observer Research Foundation 2020) ORF Issue Brief No. 389

‘Partnering to Help Curb Spread of Online Terrorist Content’ (*About Facebook*, 5 December 2016) <https://about.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content/> accessed 25 March 2021

‘Peer to Peer: Facebook Global Digital Challenge’ (*EdVenture Partners*) <https://www.edventurepartners.com/peer-to-peer-facebook-global-digital-challenge> accessed 1 April 2021

‘Pittsburgh Synagogue Shooting’, *Wikipedia* (2021) <https://en.wikipedia.org/w/index.php?title=Pittsburgh_synagogue_shooting&oldid=1010099834> accessed 11 March 2021

‘/Pol/ - On Brendon Tarrant: The Christchurch Shooter’ (15 March 2019) <https://web.archive.org/web/20190315051801/https:/8ch.net/pol/res/12919462.html> accessed 11 March 2021

‘Politicians, Gov’t Agencies Turn to Twitter amidst #Shutdown’ <https://blog.twitter.com/en_us/a/2013/politicians-govt-agencies-turn-to-twitter-amidst-shutdown.html> accessed 6 April 2021

‘Poway Synagogue Shooting’, , *Wikipedia* (2021) <https://en.wikipedia.org/w/index.php?title=Poway_synagogue_shooting&oldid=1010972146> accessed 11 March 2021

‘Product Policy Forum Minutes’ (*About Facebook*, 15 November 2018) <https://about.fb.com/news/2018/11/content-standards-forum-minutes/> accessed 1 April 2021

‘Protecting Facebook Live From Abuse and Investing in Manipulated Media Research’ (*About Facebook*, 15 May 2019) <https://about.fb.com/news/2019/05/protecting-live-from-abuse/> accessed 25 March 2021

‘Protecting Users from Government-Backed Hacking and Disinformation’ (*Google*, 26 November 2019) <https://blog.google/threat-analysis-group/protecting-users-government-backed-hacking-and-disinformation/> accessed 1 April 2021

Reed A and Ingram HJ, ‘Towards a Framework for Post-Terrorist Incident Communications Strategies’ (Royal United Services Institute for Defence and Security Studies) 12

Reed J, ‘Soldier Kills 29 People in Thailand Shooting Rampage’ (9 February 2020) <https://www.ft.com/content/8fbe1d58-4ae7-11ea-95a0-43d18ec715f5> accessed 5 April 2021

‘Reflecting on Google’s GNI Engagement’ (*Google*, 19 December 2016) <https://blog.google/outreach-initiatives/public-policy/reflecting-googles-gni-engagement/> accessed 1 April 2021

‘Removing Coordinated Inauthentic Behavior in UAE, Egypt and Saudi Arabia’ (*About Facebook*, 1 August 2019) <https://about.fb.com/news/2019/08/cib-uae-egypt-saudi-arabia/> accessed 25 March 2021

‘Report of the Australian Taskforce to Combat Terrorist and Extreme Violent Material Online’ (2019)

Reuters, ‘Netizens Circumvent Moderators to Share Christchurch Shooting Video’ (*Malaysiakini*, 08:32:00+08:00) <https://www.malaysiakini.com/news/468168> accessed 18 March 2021

Reynders D, ‘5th Evaluation of the Code of Conduct’ <https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf>

‘Royal Commission of Inquiry into the Attack on Christchurch Mosques on 15 March 2019’ (*Royal Commission of Inquiry into the Attack on Christchurch Mosques on 15 March 2019*, 2020) <https://christchurchattack.royalcommission.nz/> accessed 31 March 2021

‘Safety & Privacy on Twitter: A Guide for Victims of Harassment and Abuse’ <https://blog.twitter.com/en_us/a/2016/safety-privacy-on-twitter-a-guide-for-victims-of-harassment-and-abuse.html> accessed 6 April 2021

Sarah C Haan, ‘Facebook’s Alternative Facts’ (2019) 105 Virginia Law Review 18

School SL, ‘Making Google the Censor’ (*Stanford Law School*) <https://law.stanford.edu/publications/making-google-the-censor/> accessed 6 April 2021

‘Security Council Resolution 2354 (2017) [on Implementation of the Comprehensive International Framework to Counter Terrorist Narratives]’ <http://digitallibrary.un.org/record/1298607> accessed 6 April 2021

‘Sharing National Security Letters with the Public’ (*Google*, 13 December 2016) <https://blog.google/outreach-initiatives/public-policy/sharing-national-security-letters-public/> accessed 1 April 2021

Shead S, ‘YouTube Radicalized the Christchurch Shooter, New Zealand Report Concludes’ (*CNBC*, 8 December 2020) <https://www.cnbc.com/2020/12/08/youtube-radicalized-christchurch-shooter-new-zealand-report-finds.html> accessed 5 April 2021

‘Shedding Light on Terrorist and Extremist Content Removal’ (*RUSI*, 3 July 2019) <https://rusi.org/publication/other-publications/shedding-light-terrorist-and-extremist-content-removal> accessed 6 April 2021

‘Shitposting, Inspirational Terrorism, and the Christchurch Mosque Massacre’ (*bellingcat*, 15 March 2019) <https://www.bellingcat.com/news/rest-of-world/2019/03/15/shitposting-inspirational-terrorism-and-the-christchurch-mosque-massacre/> accessed 11 March 2021

‘Significant Progress Made on Eliminating Terrorist Content Online’ (*The Beehive*) <http://www.beehive.govt.nz/release/significant-progress-made-eliminating-terrorist-content-online> accessed 5 April 2021

‘Singapore Teenager Inspired by Christchurch Massacre Arrested for Allegedly Planning Attack on Mosques, Authorities Say | The Far Right | The Guardian’ <https://www.theguardian.com/world/2021/jan/28/singapore-teenager-inspired-by-christchurch-massacre-arrested-for-allegedly-planning-attack-on-mosques-authorities-say> accessed 13 April 2021

‘Smart Regulation for Combating Illegal Content’ (*Google*, 14 February 2019) <https://blog.google/perspectives/kent-walker/principles-evolving-technology-policy-2019/smart-regulation-combating-illegal-content/> accessed 5 April 2021

‘Social Media for Social Inclusion: Tolerance and Diversity Training’ <https://blog.twitter.com/en_us/a/2015/social-media-for-social-inclusion-tolerance-and-diversity-training.html> accessed 5 April 2021

Staff R, ‘One Gunman, Four Locations, 29 Dead: How the Mass Shooting in Thailand Unfolded’ *Reuters* (9 February 2020) <https://www.reuters.com/article/us-thailand-shooting-timeline-idUSKBN2030FQ> accessed 5 April 2021

‘Supporting New Ideas in the Fight against Hate’ (*Google*, 20 September 2017) <https://blog.google/outreach-initiatives/google-org/supporting-new-ideas-fight-against-hate/> accessed 1 April 2021

‘Supporting the Vital Work of European Safety Organizations’ (*Google*, 14 May 2019) <https://blog.google/around-the-globe/google-europe/supporting-vital-work-european-safety-organizations/> accessed 1 April 2021

‘Susan Wojcicki: My Mid-Year Update to the YouTube Community’ (*blog.youtube*) <https://blog.youtube/inside-youtube/susan-wojcicki-my-mid-year-update-youtube-community/> accessed 5 April 2021

‘Teaching Coding, Changing Lives: Google.Org Supports MolenGeek’ (*Google*, 5 June 2018) <https://blog.google/around-the-globe/google-europe/teaching-coding-changing-lives-googleorg-supports-molengeek/> accessed 31 March 2021

Team G, ‘Artificial Intelligence and Countering Violent Extremism: A Primer’ (*GNET*) <https://gnet-research.org/2020/09/28/artificial-intelligence-and-countering-violent-extremism-a-primer/> accessed 8 April 2021

‘Tech Companies Are Erasing Crucial Evidence of War Crimes’ (*Time*) <https://time.com/5798001/facebook-youtube-algorithms-extremism/> accessed 9 April 2021

‘Thai Commandos Kill Rogue Soldier Who Shot Dead 29 People’ <https://www.aljazeera.com/news/2020/2/9/thai-commandos-kill-rogue-soldier-who-shot-dead-29-people> accessed 5 April 2021

‘Thailand Shooting: Soldier Who Killed 26 in Korat Shot Dead’ *BBC News* (9 February 2020) <https://www.bbc.com/news/world-asia-51431690> accessed 5 April 2021

‘The Christchurch Attack Changed How Counter-Terrorism Thinks about Online Propaganda, Hashing and the Role of AI’ (*Faculty*) <https://faculty.ai/blog/the-christchurch-attack-changed-how-counter-terrorism-thinks-about-online-propaganda-hashing-and-the-role-of-ai/> accessed 5 April 2021

‘The EU Code of Conduct on Countering Illegal Hate Speech Online’ (*European Commission - European Commission*) <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 1 April 2021

‘The Four Rs of Responsibility, Part 1: Removing Harmful Content’ (*blog.youtube*) <https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/> accessed 1 April 2021

‘The German Synagogue Shooter’s Twitch Video Didn’t Go Viral. Here’s Why.’ <https://www.vice.com/en/article/zmjgzw/the-german-synagogue-shooters-twitch-video-didnt-go-viral-heres-why> accessed 6 April 2021

‘The Great Replacement : Featured Classification Decisions : OFLC’ <https://www.classificationoffice.govt.nz/news/featured-classification-decisions/the-great-replacement/> accessed 18 March 2021

‘The Hate-Filled Website 8chan Was Taken Offline After the El Paso Shooting. By Monday Morning It Was Back.’ <https://www.vice.com/en/article/mbmqpp/the-hate-filled-website-8chan-was-taken-offline-after-the-el-paso-shooting-by-monday-morning-it-was-back> accessed 6 April 2021

‘The Lawfare Podcast: Ben Smith on Gatekeepers in the Internet Age’ (*Lawfare*, 11 February 2021) <https://www.lawfareblog.com/lawfare-podcast-ben-smith-gatekeepers-internet-age> accessed 5 April 2021

‘The Lawfare Podcast: Canada Takes on the Proud Boys’ (*Lawfare*, 12 February 2021) <https://www.lawfareblog.com/lawfare-podcast-canada-takes-proud-boys> accessed 5 April 2021

‘The Lawfare Podcast: Content Moderation and the First Amendment for Dummies’ (*Lawfare*, 11 March 2021) <https://www.lawfareblog.com/lawfare-podcast-content-moderation-and-first-amendment-dummies> accessed 5 April 2021

‘The Lawfare Podcast: Jacob Schulz on Seditious Conspiracy’ (*Lawfare*, 24 March 2021) <https://www.lawfareblog.com/lawfare-podcast-jacob-schulz-seditious-conspiracy> accessed 5 April 2021

‘The Lawfare Podcast: Tech CEOs Head to the Hill, Again’ (*Lawfare*, 1 April 2021) <https://www.lawfareblog.com/lawfare-podcast-tech-ceos-head-hill-again> accessed 5 April 2021

‘The Lawfare Podcast: Trust, Software and Hardware’ (*Lawfare*, 22 February 2021) <https://www.lawfareblog.com/lawfare-podcast-trust-software-and-hardware> accessed 5 April 2021

‘The Lawfare Podcast: YouTube, We Have a Problem’ (*Lawfare*, 25 March 2021) <https://www.lawfareblog.com/lawfare-podcast-youtube-we-have-problem> accessed 5 April 2021

‘The New Tool Helping Asian Newsrooms Detect Fake Images’ (*Google*, 25 February 2020) <https://blog.google/around-the-globe/google-asia/new-tool-helping-asian-newsrooms-detect-fake-images/> accessed 1 April 2021

'The Redirect Method' <http://redirectmethod.org> accessed 1 April 2021

'The Report' (*Royal Commission of Inquiry into the Attack on Christchurch Mosques on 15 March 2019*) <https://christchurchattack.royalcommission.nz/the-report/> accessed 31 March 2021

'The Secret Language of Fans' <https://blog.twitter.com/en_us/topics/insights/2019/the-secret-language-of-fans.html> accessed 6 April 2021

'They Are Us' <http://shorthand.radionz.co.nz/together-alone/index.html> accessed 11 March 2021

'This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook' (*BuzzFeed News*) <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook> accessed 24 March 2021

'Three Lessons in Content Moderation from New Zealand and Other High-Profile Tragedies - Center for Democracy and Technology Three Lessons in Content Moderation from New Zealand and Other High-Profile Tragedies - Center for Democracy and Technology' <https://perma.cc/436Z-Z6J7> accessed 6 April 2021

'Tips for Engaging Live: How Automakers Used Periscope at #NYIAS' <https://blog.twitter.com/en_us/a/2016/tips-for-engaging-live-how-automakers-used-periscope-at-nyias.html> accessed 5 April 2021

'To Stop Terror Content Online, Tech Companies Need to Work Together' (*Google*, 20 December 2018) <https://blog.google/outreach-initiatives/public-policy/stop-terror-content-online-tech-companies-need-work-together/> accessed 5 April 2021

'Toomas Hendrik Ilves: Is Social Media Good or Bad For Democracy?' (*About Facebook*, 25 January 2018) <https://about.fb.com/news/2018/01/ilves-democracy/> accessed 24 March 2021

Tuesday and others, 'Shared Crisis Response Protocol | Scoop News' <https://www.scoop.co.nz/stories/PA1912/S00014/shared-crisis-response-protocol.htm> accessed 8 March 2021

——, 'Shared Crisis Response Protocol | Scoop News' <https://www.scoop.co.nz/stories/PA1912/S00014/shared-crisis-response-protocol.htm> accessed 5 April 2021

'Twitter and PBS NewsHour Partner to Live Stream Coverage of Inauguration' <https://blog.twitter.com/en_us/topics/events/2017/twitter-and-pbs-newshour-partner-to-live-stream-coverage-of-inauguration-day-2017.html> accessed 5 April 2021

'Twitter Recently Joined Young People in Vienna to Talk about Alternative Narratives & Changing Attitudes' <https://blog.twitter.com/en_us/a/2016/twitter-recently-joined-young-people-in-vienna-to-talk-about-alternative-narratives-changing.html> accessed 6 April 2021

'Twitter Supports Radicalisation Awareness Network Campaign Encouraging Europeans to #ExitHate' <https://blog.twitter.com/en_us/a/2016/twitter-supports-radicalisation-awareness-network-campaign-encouraging-europeans-to-exithate.html> accessed 6 April 2021

'Twitter's Decentralized Future' (*TechCrunch*) <https://social.techcrunch.com/2021/01/15/twitters-vision-of-decentralization-could-also-be-the-far-rights-internet-endgame/> accessed 11 March 2021

‘#UNGA: Twitter and the Global Political Conversation’ <https://blog.twitter.com/en_us/a/2013/unga-twitter-and-the-global-political-conversation.html> accessed 5 April 2021

‘Update on New Zealand’ (*About Facebook*, 19 March 2019) <https://about.fb.com/news/2019/03/update-on-new-zealand/> accessed 25 March 2021

‘Update on Our Advertising Transparency and Authenticity Efforts’ (*About Facebook*, 27 October 2017) <https://about.fb.com/news/2017/10/update-on-our-advertising-transparency-and-authenticity-efforts/> accessed 25 March 2021

‘Update on the Global Internet Forum to Counter Terrorism’ (*Google*, 4 December 2017) <https://blog.google/around-the-globe/google-europe/update-global-internet-forum-counter-terrorism/> accessed 5 April 2021

‘——’ <https://blog.twitter.com/en_us/topics/events/2017/GIFCTupdate.html> accessed 6 April 2021

‘Update on User Safety Features’ <https://blog.twitter.com/en_us/a/2015/update-on-user-safety-features.html> accessed 6 April 2021

‘US Senate Committee on the Judiciary: Opening Remarks’ <https://blog.twitter.com/en_us/topics/company/2017/opening_remarks.html> accessed 6 April 2021

‘Using Data to Change the Conversation about Race in America’ (*Google*, 13 June 2017) <https://blog.google/outreach-initiatives/google-org/using-data-change-conversation-about-race-america/> accessed 5 April 2021

Vincent J, ‘Facebook Is Now Using AI to Sort Content for Quicker Moderation’ (*The Verge*, 13 November 2020) <https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation> accessed 13 April 2021

‘Violent White Supremacy’ (*Jigsaw*) <https://jigsaw.google.com/the-current/white-supremacy/> accessed 1 April 2021

Wegener F, ‘How the Far-Right Uses Memes in Online Warfare’ (*GNET*) <https://gnet-research.org/2020/05/21/how-the-far-right-uses-memes-in-online-warfare/> accessed 13 April 2021

——, ‘The Globalisation of Right-Wing Copycat Attacks’ (*GNET*) <https://gnet-research.org/2020/03/16/the-globalisation-of-right-wing-copycat-attacks/> accessed 13 April 2021

‘What Is the Content Incident Protocol?’ (*GIFCT*) <https://gifct.org/?faqs=what-is-the-content-incident-protocol> accessed 18 March 2021

‘What to Expect on Twitter on US Inauguration Day 2021’ <https://blog.twitter.com/en_us/topics/company/2021/inauguration-2021.html> accessed 5 April 2021

‘Widows of Shuhada’ (*RNZ*) <https://www.rnz.co.nz/programmes/widows-of-shuhada> accessed 11 March 2021

Won YB, ‘Male Supremacism, Borderline Content, and Gaps in Existing Moderation Efforts’ (*GNET*) <https://gnet-research.org/2021/04/06/male-supremacism-borderline-content-and-gaps-in-existing-moderation-efforts/> accessed 8 April 2021

‘Working Together to Combat Terrorists Online’ (*Google*, 20 September 2017) <https://blog.google/outreach-initiatives/public-policy/working-together-combat-terrorists-online/> accessed 1 April 2021

‘World Leaders on Twitter: Principles & Approach’ <https://blog.twitter.com/en_us/topics/company/2019/worldleaders2019.html> accessed 5 April 2021

‘Writing Facebook’s Rulebook’ (*About Facebook*, 10 April 2019) <https://about.fb.com/news/2019/04/insidefeed-community-standards-development-process/> accessed 1 April 2021

Zannettou S and others, ‘On the Origins of Memes by Means of Fringe Web Communities’ [2018] arXiv:1805.12512 [cs] <http://arxiv.org/abs/1805.12512> accessed 6 April 2021