June 2020

Regulatory tools to address harms from content and conduct online.

A snapshot of global policy approaches.

internetnz

A research paper prepared for InternetNZ by Eloquium Group Pty Ltd.

Foreword from InternetNZ

The Internet now connects over 4 billion people and is increasingly integrated with all areas of life. With people staying at home in response to the public health emergency of the Covid-19 pandemic, those who were connected still had the potential to work and socialise, and those who were not missed out. The benefits of the Internet are now obvious to everyone.

At the same time, we are becoming increasingly aware of the new harms that people and society are facing from various types of conduct and content online. “Harms” can be defined a number of ways, but in its broadest sense could include such things as violent extremist content, the use of information about us to covertly influence what we see and how we act, the risk that people are drawn to misinformation rather than responsible journalism, unregulated advertising of harmful products to children, radicalisation of vulnerable people, lack of visibility of New Zealand culture and content, disinformation leading to public health crises or distrust in institutions, revenge porn and cyber-bullying.

These harms are real. In 2018, it was revealed that Cambridge Analytica had abused features of Facebook to scrape the personal details of tens of millions of people, and used this data to inform political campaigns and influence key voters in elections around the world. Then in March 2019, the Internet was weaponised as part of terrorist attacks on mosques in Christchurch, which were deliberately live-streamed with the aim of causing harm both to people watching, and to the social bonds of trust that knit us together.

For many people, these events were the first time they became aware of the way features of large online services could be exploited to cause real world harms to real people. People in marginalised communities were already very aware of this.

For governments around the globe, such incidents have driven and intensified efforts to identify and respond to harms people face from content and conduct online. In New Zealand the Government announced a programme of domestic law reform as part of its response to the Christchurch mosque attacks. This will include a review of media regulation which will also look at the online world.

As the New Zealand government works to understand the policy problems relating to online harms, and to make policy choices about options for addressing them, we are looking to do our part in supporting and assisting this work. InternetNZ stands for an Internet for all and an Internet for good. We are concerned about the way the Internet is being abused, but also the potential threat to the benefits of the Internet from some of the proposed responses by governments to those abuses. In working towards an Internet for good we want to tackle the harms while protecting and boosting the benefits of the Internet.

This paper is part of our contribution to the New Zealand conversation about online harms. It provides a snapshot of 20 policy options, drawn from international experience, that are available to address harmful online content and conduct. The paper does not contain a comprehensive list of potential tools, nor does it evaluate the effectiveness of the tools. It does not consider whether the tools identified are appropriate for the New Zealand context; in fact, we consider that some of them are manifestly not appropriate. The intent is that this paper will generate valuable discussion about what is appropriate.

InternetNZ looks forward to working with government, companies, community organisations and others as we all think about how to close the gap between the Internet environment we have today and the Internet environment we need. In approaching these issues, our hope is that the New Zealand government:

- takes into account the breadth, complexity and interrelated nature of the issues involved, recognises that they do not fit neatly into the remits of separate government agencies (or indeed separate governments), and works across government and internationally to develop cohesive responses;
- considers whether updating or extending existing regimes is the best approach, or whether a first principles approach including consideration of new fit-for-purpose regulatory tools would be more appropriate and effective;
- employs a transparent and inclusive process that seeks input from a diverse range of voices, draws from the vast pool of knowledge and experience available in New Zealand and elsewhere, and considers technical impacts, human rights, and the needs of New Zealand's unique culture and society.

For us the next step will be to invite and support discussion of this paper in the wider community through events and blogs and in other ways. We will also be adding to our wider ongoing programme of work towards an Internet for good with further papers and events, and we look forward to sharing those with you all as we work together to create the Internet we need.

Kim Connolly-Stone, Policy Director, InternetNZ

Jordan Carter, Chief Executive, InternetNZ

June 2020

Introduction

The Internet has been an incredible source of social and economic change, revolutionising the way Kiwis learn, entertain ourselves, explore our country, engage with business and Government services - and even the way we connect with family and friends.

With 86% of New Zealanders using the Internet every day, and 80% searching online first before buying a product or service, the Internet is becoming more important than ever for the New Zealand economy and society.[1] This has become all the more evident as the world grapples with COVID-19, seeing a shift to 'digital first' for work, learning, socialising and shopping.

Along with these social and economic benefits however, the Internet has also enabled the emergence of new forms of harmful content and conduct online, and has been used as a mechanism to amplify the impact of harmful, violent and illegal activity occurring in the offline world.

In New Zealand and around the globe, efforts to address harms from content and conduct online have increased since the Internet was weaponised in the 2019 terrorist attacks on mosques in Christchurch. New Zealand has played a leading role in increased international efforts to find meaningful and appropriate responses to these concerning developments.

As Governments grapple with online streaming of terrorist attacks and a wide range of other illegal and problematic content and conduct online such as misinformation, bullying, age inappropriate content and so on, the ability to achieve positive and meaningful change will rely on careful calibration of appropriate regulatory and non-regulatory measures. Stakeholders from across the Internet ecosystem, from individual users and policymakers, to online platforms and Internet service providers, all have a role to play in navigating these challenges.

Adding to the difficulty in finding solutions to these problems, the Internet has disrupted the way we think about regulation itself. Traditional approaches to regulating content have focused on the way users access content. For example, some traditional regulatory approaches have focused on responsibility for content, assuming that the person responsible for creating content and distributing it would be the same.

The Internet has blurred those traditional boundaries, creating a 'converged' environment in which 'content is content' regardless of the way it is accessed, or who has created it. Creation and distribution models have evolved to suit an environment where individuals can use the Internet to disseminate content across the globe.

[1] MYOB p1

Regulation needs to reflect this convergence, but also the reality that there are still important differences between many of the distribution mechanisms used to deliver content to the public, with associated differences in how best to address potential harms from a practical perspective. For example, regulatory measures that may work for an Internet business that obtains credit card information from customers can be ill-suited to address technical issues regarding access to live streamed content on a free platform.

These are complex issues and it is important to ensure that regulatory measures to address harms from content and conduct online are effective; don't create unintended consequences including in other connected areas of regulation; and don't risk undermining the positive social and economic benefits and potential of the Internet.

This paper aims to contribute to policy discussions in New Zealand about the appropriate way to achieve this when developing policy responses to address illegal, harmful and problematic content and conduct online.

Developing policy responses in New Zealand

New Zealand rightly sets itself high standards for regulation. As the Ministry for Business, Innovation and Employment (MBIE) has recognised, the quality of our regulatory systems has a major impact on the lives, work and businesses of all New Zealanders.[2] The Treasury encourages a "whole-of-system view", which involves thinking about how all the regulatory functions within a system are working together as a whole to deliver the best possible outcomes.

In addition to legislation, the Treasury has identified a number of common regulatory functions, including:

- policy advice;
- standard setting;
- operational policy/service design;
- service delivery;
- information and education;
- compliance and enforcement;
- dispute resolution; and
- monitoring and evaluation.[3]

Standards for good regulation should apply to the issues of potential harms from content and conduct online as they do to other areas.

---

2 http://regulatoryreform.com/wp-content/uploads/2015/02/New-Zealand-Best-Practice-Regulation-Model-2012.pdf

3 https://www.mbie.govt.nz/cross-government-functions/regulatory-stewardship/regulatory-systems/

What does this paper do?

As discussed, New Zealand has set expectations for the design of regulation that apply when considering how to approach regulating content and behaviour on the Internet. In implementing this consideration there are at least the following questions to be addressed:

- Whose activity is being regulated;
- Across what precise range of harms;
- Is the proposed regulatory response proportionate, and have the costs and expected effectiveness been examined; and
- Have regulatory alternatives been actively considered.

This paper focuses on the last question, asking: what are the regulatory tools that may be employed? The paper outlines a set of potential policy responses that each seek to address specific concerns in online activity.

It draws on international approaches, with examples of the design and application of each tool taken from New Zealand's peer nations. We have sourced examples from some of New Zealand's fellow Digital Nations members (Israel, Estonia, Canada, South Korea and the UK), as well as from countries with sufficiently comparable legal frameworks (France, Singapore, Ireland, Australia and Germany).

We have also included the United States as an example of an alternative approach to Internet regulation. As the home nation for many of the world's largest technology organisations, it is notable for its approach to providing safe harbours and positive incentives for Internet businesses, and minimal regulatory intervention.

We have not included examples from the efforts to apply regulation to the Internet in nations such as China, Pakistan and Russia, whose approaches in many cases would likely be inconsistent with New Zealand's Bill of Rights. However, even within the nations we have considered, there are examples of the adoption of regulatory tools that may challenge New Zealand's commitment to its citizens, or principles included in the Bill of Rights. For example, the Government of Singapore's new power to remove local access to content it deems to be 'false statements' prompted Facebook to issue a statement saying it was, "deeply concerned about the precedent this sets for the stifling of freedom of expression in Singapore."

The tools identified in the paper apply across a range of Internet policy areas, involving stakeholders at all levels of the ecosystem. In order to capture a wide range of regulatory 'tools', we have deliberately kept the scope of the paper fairly broad. We have considered harms to include not just a direct threat from criminal actors, but also misinformation, piracy and copyright infringement, electoral integrity, child safety and many other issues that may concern policymakers.

In the same vein, we have included tools that involve a range of actors. Some tools are implemented via direct legislation targeting individual users or companies, whereas others are more collaborative, or entirely led by industry, and may apply to a range of stakeholders within the ecosystem.

Some tools have been implemented, others are at an advanced level of policy consideration.

While not an exhaustive consideration of global attempts to address these issues, this paper does demonstrate a wide variety of measures for consideration, and provides real world examples where each has been exercised.

The collection of 20 regulatory tools presented in this paper forms a 'toolkit' of possible approaches and is intended as a discussion starter. Some tools identified may be appropriate for the New Zealand context, others may not.

What doesn't this paper do?

This paper provides a snapshot of a wide range of regulatory tools and examples of where they have been used, however the *effectiveness* of each tool is not discussed. Much more analysis would be required to consider the impact of the tool's use for each nation that engaged with it, how any likely positive and negative effects associated with a tool would translate into a New Zealand context and therefore whether one or more of these tools may be appropriate for use in New Zealand. Assessing whether any of these tools is effective will require a clear understanding of the policy problem it is meant to address, of how it might impact on the behaviour of online services and people using them, and of how it will operate given the global context. These questions will be an important part of the ongoing policy discussion.

It is expected that the regulatory tools outlined in this paper would be used as a *proportionate response* to clear evidence of harm. Internet NZ's paper *To block or not to block: Technical and policy considerations of Internet filtering* says, for example, that government actions affecting human rights need to be assessed against a high threshold of:

- necessity
- proportionality
- transparency
- accountability
- due process from a competent authority.

This paper does not seek to define the precise nature of any harm, or assess the suitability of any particular tool against an appropriate principled threshold for the New Zealand context. Assessing the nature of any harm will be imperative in considering whether a particular regulatory tool is a proportionate response to that harm.

Further, this paper does not outline the *costs* inherent within each tool in the form of time, effort or money. Some tools may be costly to government but have minimal impact for industry or the end user, whereas others may be relatively simple for government to implement but have a powerful disruptive effect on the technology sector - or some combination of the above. Some tools may impose a disproportionately high cost on industry to implement, when assessed against their benefits. Other tools may involve measures that may curtail economic activity, liberty or access to services for end users which would need to be carefully assessed.

It is important to keep in mind the other limitations of the approach in this paper:

- This is not a comprehensive audit of global regulatory regimes to address harms from online content and conduct. It should be seen as a 'snapshot' of various approaches overseas countries have adopted to addressing harmful online content. It is not a detailed analysis of all Internet regulation in each of the surveyed countries, nor a comprehensive summary of every country where a particular tool has been adopted.
- This paper does not seek to define the scope of organisations to which these tools might apply. This will be a critical element of any consideration of implementing a tool.
- This paper addresses regulatory approaches to harms from content and conduct online. There remains another area for exploration, which is the measures put in place to outlaw the actions of those who post harmful content or behave in a harmful way online. Countries have taken steps to impose civil and criminal penalties in certain instances. An investigation of those measures is outside the scope of this paper.

The tools in the regulatory toolbox

Tool 1	Liability limits
Tool 2	Legislated notice and takedown
Tool 3	Transparency reporting
Tool 4	Complaints and ‘trusted flagger’ systems
Tool 5	Incentivising innovation to address harmful online content
Tool 6	International collaboration and leveraging technology
Tool 7	Accreditation systems
Tool 8	Voluntary codes and industry guidelines
Tool 9	Co-regulation (Codes or industry guidelines with Regulator oversight)
Tool 10	Parliamentary and regulator scrutiny
Tool 11	Applying broadcasting standards to Internet content
Tool 12	Education and digital literacy initiatives
Tool 13	Age gating requirements
Tool 14	Administrative financial sanctions
Tool 15	Executive accountability
Tool 16	Disruption of Business
Tool 17	Filtering / ISP level blocking
Tool 18	Duty of care
Tool 19	Criminalising specific types of online content
Tool 20	No regulation/positive regulation

Opening the 'tool box'

This section of the paper provides information on each regulatory tool, and some examples of how the tool has been adopted in practice. It's important to note that depending on the context, it may be appropriate to use some combination of tools, and a number of countries have done so. Countries are listed below as examples of where a particular tool is used or proposed, and may also be using other tools or approaches not directly discussed in this paper.

Tool 1: Liability limits

What is the tool?

Some governments have introduced legal 'safe harbours' or other forms of liability limits to incentivise Internet platforms to remove harmful content from their services in exchange for limitations or exclusions from liability for acts committed by users of their services.

These approaches have been adopted partially to provide legal incentives to act, and partially to recognise the practical challenges of regulating user generated content (such as the volume of content involved, and the technical difficulties of managing content that is uploaded by users of the service rather than the service provider itself). Current New Zealand law provides broadly similar liability rules in the copyright context.[4]

Examples of the tool

United States

Section 230 of the Communications Decency Act 1996[5] (**CDA**) provides an extensive general immunity - or 'safe harbour' - from liability for publishers and users of an interactive computer service, who publish information provided by others. Section 230 provides that no 'interactive computer service' will be treated as a publisher or speaker of any information that is provided by another information content provider (which includes content posted by users of an interactive computer service).

This safe harbour means that online intermediaries that host or republish speech are protected against a range of laws that might otherwise be used to hold them legally responsible for what others say and do. Though there are important exceptions for certain criminal and intellectual property-based claims, section 230 has been claimed to be a critical element that has allowed innovation in Internet services to flourish, as well as protecting freedom of speech online.[6]

[4] See Harmful Digital Communications Act 2015, s 24; Copyright Act 1994, ss 92B & 92C; Films Videos and Publications Classification Act 1993, s 122.
[5] Title 5 Telecommunications Act 1996 (US Federal)
[6] See Electronic Frontiers Foundation https://www.eff.org/issues/cda230

Section 512(c) of the Digital Millennium Copyright Act 1996 (**DMCA**) introduced a copyright safe harbour for service providers who do not have knowledge of copyright infringements committed by users of their services, and who act expeditiously to remove or disable access to infringing content upon obtaining notice or awareness (for example, via a 'take down' notice issued by a rights holder).

Both safe harbours provide limitations of liability for service providers who comply with the conditions of the safe harbour. A critical difference between the CDA and DMCA safe harbours is with respect to immunity - legal protection is not lost under the CDA safe harbour even if the service provider fails to remove potentially illegal content once notified.[7]

Europe

The *E-Commerce Directive*[8] (2000) sought to address the issue of online service providers being liable for hosting user generated content in circumstances where active and exhaustive monitoring by such providers of their services for infringing material was a practical impossibility.

The E-Commerce Directive safe harbour provides a 'general' safe harbour that exempts intermediaries from liability for the content they host if:

- The service provider has only a neutral, technical and passive role towards the hosted content; and
- The service provider removes or disables access to the content as fast as possible once they are aware it is illegal.

There is no obligation to monitor platforms/services to obtain eligibility for this general safe harbour.

Article 17 of the 2019 Copyright Directive[9] has introduced a new form of safe harbour specific to copyright. The general E-Commerce Directive safe harbour no longer applies in relation to copyright infringements. Instead, online providers now have a copyright-specific safe harbour exempting them from liability, subject to a number of conditions. This copyright-specific safe harbour is a significant departure from the general 'no monitoring' principles set out in the E-Commerce Directive, and requires providers to have "made, in accordance with high industry standards of professional diligence, best efforts to ensure the unavailability of specific works and other subject matter for which the right holders have provided the service providers with the relevant and necessary information" (Article 17(4)).

[7] See *Zeran v America Online, Inc* 129 F.3rd 327 (4th Cir) 1997

[8] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market

[9] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market.

Australia

Similarly to New Zealand, Australia has implemented a 'safe harbour' for some types of online intermediaries that expeditiously respond to copyright 'take down' notices.[10] Intermediaries receive protection from financial damages for copyright infringements in exchange for removing infringing content from their services.

Tool 2: Legislated notice and takedown

What is the tool?

Where an online provider hosts the content they provide access to, they may operate a notice and takedown scheme. This means that where content is problematic, individuals can provide a notice which will be actioned by the provider. That action may be removing the content, asserting that content shouldn't be removed, or facilitating a counter notice process if a third party has a legitimate interest in whether the content should remain available. What constitutes problematic content can vary depending on the context.

In many contexts, this tool is implemented in collaboration with legislative immunity or 'safe harbour' (see Tool 1 above).

Examples of the tool

Australia

There are various notice and takedown schemes in Australian law, including in copyright legislation (see also discussion for Tool 1 above) and related to harmful content in the Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act (**AVM Act**). The AVM Act requires service providers to rapidly remove "abhorrent violent material expeditiously."

There are also notice and takedown schemes currently for cyberbullying content and image based abuse (also referred to as revenge porn). These latter two are currently being reviewed as part of the Online Safety Act consultation with proposals to shorten time frames for removal from 48 to 24 hours, and to extend the range of service providers to which the schemes apply.

The Online Safety Act consultation also proposes introducing a notice and takedown scheme for cyber abuse of adults. Cyber abuse of adults would be set at a higher threshold than cyber bullying, which is directed towards children.

European Union

The European Commission's proposed regulation on preventing the dissemination of terrorist content online introduces a removal order which can be issued as an administrative or judicial decision by a competent authority in a Member State. In such cases, the hosting service provider is obliged to remove the content or disable access to it within one hour. Systematic failure to meet this time frame may be sanctioned with a fine of up to 4% of the respective provider's global turnover.

[10] See Part V Division 2AA *Copyright Act 1968*.

The European Parliament voted for a comprehensive overhaul of the proposal, including explicitly excluding from scope non-public content hosting such as messaging and cloud infrastructure services and limiting authority to issue takedown notices to entities with functionally independent administrative authority, to prevent government censorship.

The legislation is now in 'trilogue' with the Council, Parliament and Commission.

France

The French Parliament passed a new law in October 2018 which allows candidates and political parties to appeal to a judge to help stop false information during the three months before an election. A 'quick response' process is established to enable a court to rule on whether reports published are credible or should be taken down (within 48 hours). Violating the law is punishable by up to one year in prison and a fine of €75,000. The French Audiovisual Council, the broadcasting body that regulates radio and television, is also empowered to block foreign state-controlled broadcasters that publish false information.

A far more widespread regulatory effort is also underway in France. The proposed 'Avia Bill' (Project de loi Avia) passed the French National Assembly in mid-2019, while the Senate passed an amended version this year. The Joint Parliamentary Committee has so far been unable to reach a compromise between the two versions, meaning that at the time of writing the Bill is not yet in effect.[11]

If passed, the Bill would require communication service providers to remove offending content within 24 hours of receiving notice. The Bill targets texts, pictures, videos and web pages that incite hatred or violence, or that carry insults of a racist or religious nature. The Bill would also require communication service providers to comply with a number of additional requirements, including:

- Companies must put in place internal complaints mechanisms and provide information about external avenues of appeal
- The mechanism for reporting offending content must place no requirement on reporting individuals to justify why they believe the content to be illegal
- Companies must have a legal representative within the country in which they operate
- Fines of up to 4% of annual turnover for serious and recurrent failures to remove
- Companies must allow a specialist law-enforcement body to order blocking or de-referencing of websites, servers or electronic access to content deemed illegal by court decision
- Fines for failing to preserve data that might identify offending users of up to €250,000, triple the current fine of €75,000.

[11] Following the final draft of this paper, the French Constitutional Court reversed most of the provisions of the Avia Law passed by French National Assembly, on the basis that the provisions would have been in violation of the French Constitution. The Court noted the turn around times (24 hours for manifestly unlawful content and 1 hour for terrorist and child sexual abuse content) would likely to lead to overblocking content online with potentially discriminatory effects. See *Jurist* "France Constitutional Court strikes down most of online hate speech law" (June 20, 2020) <jurist.org>

There is significant concern from the EU over this Bill, suggesting that the short and non-negotiable timeframes will lead to over-censorship and filtering of content, and may conflict with the European Union's e-Commerce Directive. The European Commission wrote to Minister Le Drian:

> "The Commission shares with the French authorities the policy objective of fighting illegal content online. However, in view of the Commission's intention and on-going work towards proposing and adopting EU legislation on the matter in the near future, it is suggested that Member States exercise restraint and postpone the adoption of national initiatives on this same matter, such as the notified draft." [12]

Germany

Through the Netzwerkdurchsetzungsgesetz (Network Enforcement Act or NetzDG), Germany requires online service providers to have reporting mechanisms and remove manifestly unlawful posts within 24 hours. Content is considered unlawful if it is illegal under the German Criminal Code. Fines apply for failure to remove. The law was implemented in 2018 and has reportedly led to Facebook hiring German speaking content moderators.

Tool 3: Transparency reporting

What is the tool?

Opacity around the operations of an organisation, or the technologies used in the delivery of Internet services, can make it difficult for the public and policy makers to ascertain whether the organisation is upholding appropriate standards. Publication of transparency reports can provide insight into the operations of organisations. Transparency about how technology operates 'under the hood' can provide comfort to regulators about implementation of policies and standards.

In some instances, organisations will voluntarily elect to publish transparency reports. In other situations governments have required transparency reports (such as the publication of key metrics and data) from a select group of organisations for which such insight is deemed necessary, or engaged directly with companies to better understand business and technical operations.

Examples of the tool

International

Transparency is one of the commitments included in the Christchurch Call (see below at Tool 6). In this instance, transparency reporting is limited to terrorist and violent extremist content that is detected and removed by online service providers. It is intended that reporting will be measurable and supported by clear methodology.

[12] http://ecnl.org/wp-content/uploads/2020/01/EU-Commission-Opinion-Avia-Bill-draft.pdf

The Global Internet Forum to Counter Terrorism (**GIFCT**) (see below at Tool 6) is also publishing transparency reports. Its first report, published in July 2019 is available here.

In addition, the OECD is working on developing a standardised transparency report protocol, so that social media platforms can all rely on a common, simple, internationally harmonized reporting mechanism on terrorist content.

United Kingdom

The UK Government announced in its response to the Online Harms White Paper Consultation, that it plans to implement transparency reporting in an effort to cultivate a culture of transparency, trust and accountability.

The UK approach is designed to ensure the regulator can gain an understanding of the harms occurring on online platforms, and the action being taken by companies in response. Transparency will also enable consumers to better understand which companies are taking positive steps to keep their users safe, and the processes that different companies have in place to prevent harms.

The planned approach will see annual transparency reports required, outlining the prevalence of harmful content on their platforms and what counter measures companies are taking to address these. It is also possible that the regulator will be empowered to request additional information.

Throughout the consultation process there was concern expressed about a potential 'one size fits all' approach to transparency, and the material costs for companies of reporting. The Government has announced that reporting requirements will vary in proportion with the type of service being provided and the risk factors involved. As such, the regulator will apply minimum thresholds in determining the level of detail a company needs to provide, or indeed whether it needs to provide a transparency report at all.

Ireland

The proposed Online Safety Commissioner will be given the power to create rules about periodic reporting of compliance with online safety codes by regulated online services.

Australia

Transparency reporting is being considered in Australia through the Online Safety Act consultation process. The Government intends to implement requirements to provide Transparency Reports that "provide data on the number and type of responses to reports and complaints about illegal, abusive and predatory content by users."

Under the proposal, the eSafety Commissioner would have the power to determine that particular organisations must provide a report. Organisations would be selected based on criteria such as numbers of complaints received, size of user base and significance of harmful activity. It is proposed that there would be penalties for non-compliance.

The Government is aware of the potential regulatory burden imposed on industry, as well as international moves towards transparency reporting. As such, it is proposing to establish a reporting framework that would integrate with other efforts, including the OECD's voluntary transparency reporting protocol currently being developed.

The reporting framework would also be designed in a way that meets the objectives of the Taskforce to Combat Terrorist and Extreme Violent Material Online (also referred to at Tool 17). The Report of that Taskforce determined that providers should publish reports outlining their efforts "to detect and remove terrorist and extreme violent material on their services". This could extend to how pieces of content are identified and whether they were engaged with before removal.

France

In France's work to embed civil servants within Facebook (see below at Rule 10), the resulting report proposes transparency obligations on social networks relating to their ordering of content and how terms of service are enforced.

Europe

The European Union are among policy makers who have been scrutinising the issues arising from the practices of businesses that increasingly use algorithms to automate many of their processes. This can be from using AI driven 'chat bots' to respond to customer queries via websites, to managing display advertising, "hash"-based content and copyright management systems, to the algorithms underpinning the content that is surfaced to users on social media platforms.

'Algorithmic transparency' can also be characterised as a desire to 'peek under the hood' to see the technological processes used by an Internet business to understand whether there may be data privacy, competition or consumer protection concerns about the way businesses operate.

The European Union released a research paper setting out a possible governance framework for algorithmic accountability and transparency. Four policy options are proposed, addressing an aspect of algorithmic transparency and accountability:

- awareness raising: education, watchdogs and whistleblowers;
- accountability in public-sector use of algorithmic decision-making;
- regulatory oversight and legal liability; and
- global coordination for algorithmic governance.

Tool 4: Complaints and 'trusted flagger' systems

What is the tool?

Customer support and the ability to make complaints is commonly available in traditional industry, and has historically been less available from online service providers who operate at vast scale. In contrast, major online platforms have invested millions of dollars in 'flagging' or 'report abuse' systems to enable users

to report or 'flag' inappropriate or harmful online content. NetSafe provides information about the reporting systems for major online platforms and additional support for New Zealanders.

There have been moves to impose additional complaints and response obligations upon online providers, or in some circumstances for regulators or trusted non-government organisations to have enhanced status as a 'trusted complainant' or 'trusted flagger'. This is familiar in the New Zealand context, with NetSafe operating as a 'trusted flagger' for certain online providers.

Examples of the tool

Australia
Over the past two years, the ACCC has been examining digital platforms, in particular "the effect that digital search engines, social media platforms and other digital content aggregation platforms have on competition in media and advertising services markets."

In its Final Report, the ACCC recommended that the ACMA develop minimum internal dispute resolution standards that must be adopted by digital platforms, as well as the establishment of an ombudsman scheme to resolve complaints and disputes with digital platform providers.

In its response, the Government committed to work with major digital platforms to scope and implement a pilot of an external dispute resolution mechanism for complaints between consumers, businesses and digital platforms. The Government will assess the development and rollout of the pilot scheme over the course of 2020, along with any parallel improvements in associated internal dispute resolution processes. This will inform the Government's consideration during 2021 of the need for a broader external dispute resolution process, including a Digital Platforms Ombudsman.

United Kingdom
Following its Online Harms White Paper consultation, the UK Government has committed to implement regulator (likely OFCOM) oversight of companies' complaints processes. The regulator will receive transparency information about the volume and outcome of complaints and have the power to require improvements as necessary. It will not receive or make decisions on specific complaints.

Under the duty of care the UK plans to implement (see below), the codes that set out how companies will be expected to implement that duty, are likely to include that companies should have effective and easy to use complaints functions.

Ireland
Under the proposed Irish online safety scheme, the Online Safety Commissioner will operate a "super complaints" scheme for nominated bodies (such as expert charities) to bring issues with online services to the Commissioner's attention. This is similar to a 'trusted flagger' system operated by a digital platform, but implemented by the regulator to help manage complaints.

The Online Safety Commissioner will establish a scheme to receive "super complaints" about systemic issues with online services from nominated bodies, including expert NGOs, and may request information, investigate or audit an online service on the basis of information received through this scheme.

Israel
The Cyber Division of the Israeli Justice Ministry undertakes proactive monitoring of online content, including social media posts, and contacts companies such as Facebook and Twitter to secure removal of posts that violate Israeli law, or the terms of service of the platforms hosting the content.[13]

Tool 5: Incentivising innovation to address harmful online content

What is the tool?

Another regulatory tool is to encourage technological measures to assist in addressing harmful online content. In other words, if use of technology in a particular way has caused the problem, can we encourage the design of that technology in a way that solves for that use?

This tool could be included as part of industry initiatives, or specifically encouraged or required by governments as part of formulating policy responses to addressing harmful online content.

Examples of the tool

Ireland
The Irish Government has announced plans for a new Online Safety Act. Although at the time of writing the detail of the legislation was still being prepared, the Government has announced the 'heads' of what will be in the legislation. Online platforms will be required to implement appropriate technical and organisational measures to minimise the risk of online harm, and to "regularly review and update when necessary" the technical and organisational measures implemented. This obligation is expected to include an obligation to phase Artificial Intelligence (AI) initiatives into the technologies used by digital platforms over time to continue to innovate in improving safety.[14]

[13] See https://www.middleeasteye.net/fr/news/israel-moves-censor-online-content-violates-israeli-law-1075211112

[14] McCann Fitzgerald analysis General Scheme of the Online Safety & Media Regulation Bill 2019 published, ending the era of self-regulation for online media in Ireland, 10 January 2020 https://www.mccannfitzgerald.com/knowledge/media-and-entertainment/general-scheme-online-safety-media-regulation-bill-2019-published

Including obligations regarding technological innovation and 'continuous improvement' into the safety requirements imposed on digital platforms reflects a recognition of the importance of continued innovation in the technical measures used to protect safety, and that what is considered to be reasonable with regard to deploying technical measures may evolve over time.

United Kingdom

As discussed in more detail at Tool 18, the UK Government plans to implement a duty of care regarding online harms, supported by industry codes. One of the issues identified for potential inclusion in industry codes is an obligation on providers to regularly review their efforts to tackle harm and drive continuous improvement, which would include technological improvement, to ensure services are safe by design.

Tool 6: International collaboration and leveraging technology

What is the tool?

The global nature of many online services means that one government will not necessarily be able to impact the operations of a service across the globe. Rather than act in a piecemeal fashion, there are increasing efforts to coordinate international action and work collaboratively with service providers to promote appropriate use, behavior and content on online services.

Examples of the tool

Christchurch Call and the Global Internet Forum to Counter Terrorism

The Christchurch Call is a commitment by Governments and tech companies to eliminate terrorist and violent extremist content online. It sets out collective, voluntary commitments from Governments and online service providers to address this content and prevent abuse of the Internet in the way that happened in Christchurch in March 2019. There are a set of commitments for Governments, and a separate set for online service providers.

The Global Internet Forum to Counter Terrorism (**GIFCT**) is dedicated to preventing terrorists and violent extremists from exploiting digital platforms. It was established in July 2017 as a group of companies with a rotating chair drawn from the member companies, it was focussed on knowledge sharing, technical collaboration and shared research.

Following the Christchurch Call, the GIFCT evolved into an independent organisation which aims to sustain and deepen industry collaboration and capacity while incorporating the advice of key civil society and government stakeholders. It now has ongoing dedicated staff and works on prevention, response and shared research.

European Union Code of Conduct for combating hate speech online

Facebook, Twitter, Microsoft, and YouTube agreed in 2016 to a new Code of Conduct that requires them to review "the majority of" hateful online content within 24 hours of being notified, and to remove it, if necessary, in the name of combating hate speech and terrorist propaganda across the EU. The Code of

Conduct puts more responsibility in platforms to police content, without the accountability and oversight of democratic institutions.

Project Arachnid

Project Arachnid is a project to trawl the web to identify web pages with suspected child sexual abuse material. The technology can be deployed across websites, forums, chat services and newsgroups to detect illegal content, before sending a takedown notice to service providers so they can quickly act.

The project is led by Canada with support from a range of countries and organisations including the UK government, Australia's eSafety Commissioner and the US through the National Centre for Missing and Exploited Children. As at March 2019, Arachnid had trawled 1.5 billion webpages, detected 7.5 million suspected images of child sexual abuse and issued more than 1 million take-down notices for the removal of child abuse material on the open web.

Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse

In March 2020, the US Attorney General published the Voluntary Principles, a set of 11 principles that companies in the technology industry can choose to adopt in order to protect children from online predators. The principles were developed by the Five Eyes (FVEY) nations in consultation with leading technology companies - Twitter, Facebook, Google, Microsoft, Roblox and Snap.

The principles encourage companies to take an active role in identifying and preventing child sexual exploitation activity, preventing search results and dissemination of exploitative material, report exploitative material to an appropriate authority and engage in knowledge sharing to raise standards across industry.

The principles will act as a common framework for organisations to assess their own safety standards and processes, identify gaps in their systems, understand the level and nature of online child sexual exploitation and respond to the evolving threat in order to reduce risks for users.

Tool 7: Accreditation systems

What is the tool?

With such a wide variety of information and services available online, there are instances where countries have explored establishing accreditation systems to 'accredit' certain services or programs. The intention is to have accredited services/programs display a trustmark or similar, so that users know a certain standard has been ascertained.

There are limitations on implementation of this approach, notably capacity to continually refresh assessment of multiple services which regularly change, and the implications for services/programs not accredited because of being out of scope or because of resource constraints, rather than because of falling foul of requirements.

Examples of the tool

Australia
The eSafety Commissioner runs the Trusted eSafety Provider Program. Under the program, "individuals and organisations delivering high quality online safety education programs in schools can apply to become a Trusted eSafety Provider". This helps schools and the community choose a high quality provider of education services.

The Online Safety Act Consultation explores setting up an accreditation system for safety tools and services, led either by the eSafety Commissioner or industry. This is in response to "the diversity of tools and services (tending to confuse) users trying to determine the best way of protecting children and other vulnerable people from inappropriate material."

Canada
Canada's wide-ranging review of its broadcasting and telecommunications regulatory framework Canada's Communications Future has recommended that the existing licensing regime in the Broadcasting Act be accompanied by a registration regime that would require a person carrying on a media content undertaking via the Internet to register. (Recommendation 56).

Tool 8: Voluntary codes and industry guidelines

What is the tool?

For many years, the online world has been the purview of self regulation - efforts by key organisations to set standards for their operation that they self-enforce and to which they are held to account, by the sentiment of their users, the wider public and key stakeholders such as governments. These standards are in some instances convened by an industry body and in other instances are done through collaboration of key players.

The tools identified here focus on voluntary initiatives from industry, sometimes in partnership with, or at the request of governments or regulators.

Examples of the tool

Europe
The EU was an early mover in working with industry to provide guidance on best practice to prevent online harms. The Safer Social Networking Principles were established in 2009. These saw signatory companies adhering to a range of Principles including user education, provision of user tools to tailor their experience, reporting mechanisms and promotion of compliance with terms of service. Each signatory provided a Self Declaration Form, outlining how they would meet the requirements of the Principles. The Principles also set up a cadence for industry discussions about trends and developments.

More recently, the European Commission has developed a Code of Conduct on Countering Illegal Hate Speech Online. Since May 2016, Facebook, Twitter, YouTube and Microsoft have committed to combatting the spread of racist and xenophobic content and terrorist propaganda in Europe through this code. Other platforms more recently announced they will participate under the Code, including Instagram (January 2018), Snapchat (May 2018), Dailymotion (June 2018) and Jeuxvideo.com (January 2019). The last evaluation shows that this Commission initiative delivers successful results: the companies are now assessing 89% of flagged content within 24 hours and 72% of the content deemed illegal hate speech is removed.

Australia

In Australia, there have been various efforts to put in place industry guidelines and voluntary codes.

For example, the eSafety Commissioner is leading work on Safety by Design principles. These are "guidelines that provide a model to assess, review and embed user safety into online services". They provide a benchmark for industry, against which to assess, review and embed user safety into the design of online services. The core Safety by Design principles were established in early 2019, and work is currently underway to create a Framework of guidance for industry use.

As an earlier example, in 2014, those involved in digital advertising established the Online Behavioural Advertising Guidelines. These guidelines set out parameters for notice, choice and accountability in the provision of interest based advertising, to which providers would adhere. Oversight was by the Australian Digital Advertisers Association (ADAA). In following years, user and stakeholder views created momentum in the industry towards the provision of ad blockers, allowing users to block certain types of ads from their browsers. The industry's desire to provide useful, relevant ad experiences underpins continued efforts to provide users with control.

Canada

Charters can function as guidelines to industry in terms of what is broadly expected of them. The Canadian Government, after consultation with industry, launched a Digital Charter in 2019 that looks to foster an "innovative, people-centred and inclusive digital and data economy". The charter consists of ten principles to this end, including Universal Access, Safety and Security, Free from Hate and Violent Extremism and Strong Enforcement and Real Accountability.

In terms of industry-led initiatives, Canada's Public Policy Forum, an independent non-profit think tank for public-private dialogue, has proposed a Moderation Standards Council, analogous to the Canadian Broadcast Standards Council, adapted for online content. The proposed Council would comprise stakeholders from across government and industry, acting to help online content providers meet public expectations and government requirements. It would function as a resource for code of conduct development, facilitate appeals processes on content moderation decisions and address jurisdictional conflicts.

Tool 9: Co-regulation (Codes or industry guidelines with regulator oversight)

What is the tool?

In many ways an evolution of voluntary or industry self regulation, co-regulation involves industry and government jointly administering standards. For example, industry may set codes and guidelines outlining standards to which organisations will adhere, and a regulator may then oversee and in some instances enforce compliance with those codes and guidelines. There may even be a power for intervention where satisfactory codes and guidelines are not put in place.

Examples of the tool

Australia

Australia currently has a system of co-regulatory codes to address harmful content and conduct online.

An objective of Australia's broadcasting legislation is to enable "public interest considerations to be addressed in a way that does not impose unnecessary financial and administrative burdens on industry".[15] This has led to a preference for a co-regulatory approach in the sector.

For example, industry codes are established under Schedules 5 and 7 of the Broadcasting Services Act (**BSA**) to address harmful online content. Together, the scheme places constraints on the types of online content that can be hosted or provided by Internet service providers and content service providers. Schedule 5 governs Internet content hosted outside Australia, and Schedule 7 addresses content services provided in Australia, including some content available on the Internet. More discussion of the nature of the age gating obligations in the codes is provided below at Tool 13.

The BSA specifies the matters that must be dealt with by industry codes, and in some cases establishes additional matters that may be included in codes. For example, regarding offshore content, the code must address measures for enabling parents to better monitor the online activities of their children, provision of filtering technologies, content labelling, legal assessments of content, and complaints handling procedures.[16]

The Australian Communications and Media Authority (**ACMA**) investigates complaints relevant to the codes and takes appropriate action to enforce if necessary.

There is currently consultation underway on a new Online Safety Act, which includes a proposal to update this codes process. The consultation paper states that Australia "would retain the provisions for online service providers to develop codes of practice to address harmful online content and for the eSafety Commissioner to make an industry standard should these codes prove to be

[15] Subsections 4(2)(a) and 4(3)(a) *Broadcasting Services Act 1992.*
[16] Schedule 5 clause 60 BSA.

ineffective. The eSafety Commissioner would retain powers to refer sufficiently serious content to law enforcement for investigation. However, the code provisions would be updated to require codes to be principles based and stipulate that codes should be developed by a wider range of service providers than the current codes, reflecting the range of online services that Australians now use to access online content."

United Kingdom

The UK plans to implement codes with regulator oversight, most likely the Office of Communications (OFCOM). Its Response to the Online Harms White Paper Consultation affirms that it will proceed with a duty of care supported by codes (see below), and will immediately work to put in place codes of conduct regarding terrorist activity or child sexual exploitation or abuse (CSEA). The UK Government plans to publish codes of practice that provide guidance for tackling these online harms with companies being required to take robust action.

The types of inclusions we could expect in the UK codes regarding terrorist activity and CSEA include requirements that companies:

- Ensure their terms and conditions meet standards set by the regulator; are readily understood and are enforced effectively and consistently.
- Take reasonable steps to prevent known terrorist of CSEA content being made available to users.
- Respond promptly, transparently and effectively to user reports; and provide appeal processes for the removal of content or other responses.
- Support law enforcement investigations.
- Direct users who have suffered harm to support.
- Regularly review their efforts to tackle harm and drive continuous improvement; and more generally take reasonable steps to ensure services are safe by design.

Ireland

Ireland will establish an Online Safety Commissioner to support its new Online Safety and Media Regulation Act. It is intended that the regulator will set rules and norms and establish the content of what should be in the codes that apply to various sectors of industry. The codes will apply to a wide range of matters, including harmful online content, commercial communications, risk and impact assessments, and complaints handling.

France

In France's work to embed civil servants within Facebook (see below at Tool 10), the resulting report proposed a co regulatory approach. In particular, it proposed "co-regulatory mechanisms that impose the internal assimilation of public interest objectives, without defining the methods."

Israel

The Israeli government and Facebook agreed in 2016 to work directly together to determine how to tackle incitement on the social media network. Israeli security agencies are reported to monitor the service for 'incitement' and complain to Facebook, to enable the company to determine whether the content violates Facebook's community standards.

Tool 10: Parliamentary and regulator scrutiny

What is the tool?

It is common for governments to scrutinise the operations of online providers both through Parliamentary and regulator scrutiny. This can involve using powers to establish Parliamentary reviews and regulator investigations of certain issues as a regulatory tool for addressing harmful online content.

Examples of the tool

Australia

In Australia there have been multiple Parliamentary inquiries into issues arising in relation to online service providers. These include the Select Committee on Foreign Interference through Social Media and the Select Committee on the Future of Public Interest Journalism.

Alongside Parliamentary scrutiny there has been regulator scrutiny. This includes investigations by the Privacy Commissioner, for example in relation to Facebook's handling of disclosure of information to Cambridge Analytica.

The ACCC has also been examining digital platforms since late 2017, in particular "the effect that digital search engines, social media platforms and other digital content aggregation platforms have on competition in media and advertising services markets." An 18 month long investigation culminated in a series of recommendations to which the Government responded in late 2019.

The Government response included funding for the ACCC to continue to examine competition and consumer protection issues involving digital platforms. In the 2019–20 Mid-Year Economic and Fiscal Outlook the Government is committing $27 million over four years to establish a Digital Platforms Branch within the ACCC. The new Branch is empowered to:

- monitor and biannually report on digital platforms;
- take enforcement action as necessary; and
- conduct specific inquiries as directed by the Treasurer, the first of which is an inquiry into competition for the supply of ad tech services and the supply of online advertising by advertising and media agencies.

Another aspect of Government scrutiny of online provider activity is the Taskforce to Combate Terrorism and Extreme Violent Material Online. This Taskforce was set up in the wake of the 2019 terrorist attacks on mosques in Christchurch and brings together relevant Government agencies, Internet service providers and online providers. It's focus was to unpack the response to Christchurch and advise on "practical, tangible and effective measures and commitments to combat the upload and dissemination of terrorist and extreme violent material."

European Union

The Executive Vice-President of the European Commission, Margrethe Vestager, has assumed responsibility for setting the strategic direction of the political priority, Europe Fit for the Digital Age. As part of her ongoing regulatory scrutiny, her office has made a number of high profile investigations into large technology companies and their operations, including:

- a formal inquiry of potential exclusionary practices by technology firm Broadcom.
- an inquiry over whether Amazon unfairly uses data collected from third party sellers who rely on its platform.
- Facilitating an antitrust inquiry into Apple after it was accused by music streaming service Spotify of anti-competitive behaviour.
- Ruling against Google for abusing its dominance with Adsense advertising service and its mobile operating system, Android, and fining the organisation €1.49 billion for abusive practices in online advertising.

France
Following discussions in late 2018, the Facebook CEO and the French President announced a team of civil servants would spend time embedded in the company. Their role was to work in conjunction with Facebook to determine concrete, tailored proposals to fight hate speech, while monitoring Facebook's policies and tools for stopping posts and photos that attack people on discriminatory grounds. It resulted in a report which states that self regulation is not adequate to respond to the potential harms associated with social networks. Instead the report recommends the need for co-regulation (see above at Tool 9).

Israel
The Israeli government and Facebook agreed in 2016 to work directly together to determine how to tackle incitement on the social media network. It has been reported that the Interior Minister's Office agreed with Facebook to create teams to agree to remove "inflammatory content".

Tool 11: Applying broadcasting standards to Internet content

What is the tool?

From time to time there is discussion of including online service providers in the scope of broadcasting legislation which would see them subject to requirements to classify content, and to comply with other broadcasting standards regarding fairness and accuracy of news and current affairs, 'decency' standards, standards regarding the broadcast of particular types of content, local content commitments, captioning obligations and similar broadcast-related approaches. In practice however, there are significant differences in the nature of operation of broadcasters and online service providers which makes direct translation of the legislation problematic.

Examples of the tool

European Union
The *Audiovisual Media Services Directive* (**AVMSD**) is a legislative act of the EU that creates a framework for cross-border audiovisual media services. The first version of the Directive was created 28 years ago to standardise regulation for television programmes and allow broadcasting between member states.

The framework underwent a review from 2016-2018 with the purpose of unifying standards across online content, creating a Single Digital Market. In its 2016 iteration, video streaming services such as Netflix were the only digital providers

covered under the AVMSD. The revised AVMSD applies to video sharing platforms such as Youtube and the video content on social media services such as Facebook.

The AVMSD provides various obligations for media content providers, including requiring appropriate measures for protecting children from harmful content, protection from incitement to violence or hatred, protection for children from inappropriate commercial communications for unhealthy foods, transparency about commercial communications and a requirement for audiovisual regulators in member states to be legally distinct from their government.

Finland, Ireland and the Netherlands voted against the inclusion of video sharing platforms and social media under the revision of the AMSD, saying: "*as stated consistently during the negotiations, the AVMS Directive is not the correct place for regulating video sharing platforms since the rest of the scope of the directive covers only AV media services where the service provider has editorial responsibility for the content of the program.*"

Canada

Canada's wide-ranging review of its broadcasting and telecommunications regulatory framework Canada's Communications Future has recommended the scope of the Broadcasting Act extend to alphanumeric news content (Recommendation 51) and that the Broadcasting Act apply to 'media content undertaking', which would replace the term 'broadcasting undertaking' in the Act (Recommendation 54).

Australia

Australia decided to expressly exclude Internet content from its broadcasting legislation via a Ministerial Determination made in 2000, and extended again in 2019. This has meant that emerging Internet services, including user generated content services such as YouTube and streaming services such as Netflix, have been excluded from the majority of provisions in the Broadcasting Services Act.

The ACCC Digital Platforms Report highlighted the differential application of various regulatory standards in the media and communications regulatory landscape, and recommended that the Government commence a policy review process to develop a harmonised media framework.

> ***Recommendation 6***
> *A new platform-neutral regulatory framework be developed and implemented to ensure effective and consistent regulatory oversight of all entities involved in content production or delivery in Australia, including media businesses, publishers, broadcasters and digital platforms. This would create a level playing field that promotes competition in Australian media and advertising markets.*

The Government has accepted this recommendation, and a review is expected to be commenced in 2020.

Tool 12: Education and digital literacy initiatives

What is the tool?

Digital literacy, or capability, is essential for social and economic participation in the digital world. It is also essential to safely navigate online risks and potential harms. Being digitally literate, or digitally capable, means an individual can safely and securely operate devices, communicate and socialise online, access and think critically about information online, transact online, problem solve using online tools and around issues arising with technology. At its best, it extends to supporting growth mindsets and lifelong learning as technology develops. The importance of digital literacy initiatives has been highlighted by the COVID-19 crisis, where capability for all New Zealanders is vital to accessing essential information, distinguishing valid information, maintaining work, participating in education and for social connection.

There is an array of programs and efforts to support citizens' development of digital literacy. In some instances, these programs and efforts are coordinated through an overarching digital literacy/capability strategy. There are also examples of countries having 'Digital Capabilities Frameworks' which provide a common understanding of the skills/capabilities each citizen should be encouraged and supported to develop the skills to protect themselves from harmful online content and conduct.

In some situations, education initiatives have been adopted with government officials and regulators, to ensure policy makers have sufficient skills to make nuanced and beneficial policy decisions. Digital literacy and technological capability is essential if policy makers are to be effective in understanding the challenges and developing effective responses.

Examples of the tool

Australia

Australia has a variety of initiatives underway to support the development of digital literacy. These are the efforts of Government, community organisations and private sectors. Within the private sector, these include: Go Digi (for individuals), Digital Springboard (for individuals), Digital Garage (for small businesses) and Tech Savvy Seniors. There are also multiple Government Departments working on digital inclusion programs including Be Connected (directed towards older seniors), Get Online Qld, Digital Ready Tas and resources from the eSafety Commissioner.

These programs of work operate independently of one another and are not coordinated through an overarching digital literacy strategy.

United Kingdom

The UK is currently working on a new online media literacy strategy. It seeks to ensure a coordinated and strategic approach to online media literacy education and awareness for children, young people and adults. It is expected to be published in the summer of 2020.

As part of the 2017 UK Digital Strategy, the government established the Digital Skills Partnership (DSP), which brings together public, private and charity sector organisations to determine best practice in the digital skills environment. The DSP has three priorities:

- Supporting the development of Local Digital Skills partnerships in English regions.
- Increasing digital enterprise by helping small businesses and charities upskill.
- Support computing in schools.

The UK's Department for Education publishes the Essential Digital Skills Framework. The framework is intended to be used by everyone in the UK engaged in supporting adults to enhance their digital skills, including an annual measurement of the framework in the form of the Lloyds Bank Consumer Index.

Canada

There is no overarching digital literary strategy from the federal government, contrasting with, for example, a significant investment in financial literacy, where the government convened a task force in 2009 and created the Financial Consumer Agency of Canada.

Currently, digital literacy education and resources are provided by industry and not-for-profits such as Mediasmarts, Canada's centre for digital and media literacy, which provides educational resources to improve user detection of disinformation and misinformation online.

Expanded policy priority and funding for digital literacy as part of a co-ordinated federal government program is recommended in Democracy Under Threat (Recommendation 17) and Canada's Communications Future (Recommendation 87).

Estonia

Estonia has included digital literacy and capabilities for its public sector into its Digital Agenda strategy, including increasing the capability of its public sector to use data analytics and research.

Tool 13: Age gating requirements

What is the tool?

In certain circumstances, laws may require that online providers verify a user's age before granting access to certain content. The intent is to restrict access to only those of an appropriate age. There are significant limitations on the effectiveness of age verification, including how age can effectively be ascertained in a way that is sensitive to privacy considerations.

Examples of the tool

Australia
Schedule 7 of the Broadcasting Services Act prohibits content service providers from making available "prohibited" or "potentially prohibited content". Prohibited content is content that has been classified by the Classification Board as illegal or legally restricted (Refused Classification or X18+), or in some cases, content that has been classified as age restricted (R18+ or MA 15+) and that content is not subject to a "restricted access system". (Content is 'potentially prohibited' if it has not been rated by the Board, but there is a substantial likelihood that it would be found to be prohibited content if it were to be rated).

The technical requirements for a Restricted Access System are set out in regulations, but in general a system must implement reasonable steps to verify that a user is 18 or above, or in relation to MA15+ content, that the system is capable of verifying that the applicant has declared they are over 15 and sufficient warning information about the content has been provided.

A recent report from the House of Representatives Standing Committee on Social Policy and Legal Affairs considered the potential for a wider online age verification regime to protect children and young people in Australia from exposure to online wagering and online pornography.

The Committee recognised that age verification is not a 'silver bullet', and that protecting children and young people from online harms requires government, industry, and the community to work together across a range of fronts. However, the Committee also concluded that age verification can create a significant barrier to prevent young people—and particularly young children—from exposure to harmful online content, and recommended that online age verification be implemented in Australia.

The Committee recommended that the Digital Transformation Agency lead the development of standards for online age verification, to ensure that online age verification is accurate and effective, and that the process for legitimate consumers is easy, safe, and secure. The Committee also recommended that the Digital Transformation Agency develop an age-verification exchange to support a competitive ecosystem for third-party age verification in Australia.

United Kingdom
In contrast to the Australian Parliamentary Committee's recommendation to adopt an age verification system, the UK has recently abandoned its plans to implement age verification for online adult content. Instead, the government will focus on measures to protect children in the much broader online harms white paper initiatives, which are discussed in more detail in the context of tool 18 below.

South Korea
The *Youth Protection Revision Act 2011*, (commonly known as the 'Shutdown Law' or 'Cinderella Law'), is an act of the South Korean National Assembly that forbids children under the age of sixteen to play online video games between the hours of 00:00 and 06:00. Industry was required to develop technical solutions to age gate access to services to enforce the laws. Most providers have used the national

social security number to underpin the technical capacity to implement this law, with most services choosing to 'lock out' those users without a social security number during the shutdown time.

Tool 14: Administrative financial sanctions

What is the tool?

In most countries that recognise a version of the Westminster doctrine of 'separation of powers', there are constitutional challenges to imposing civil fines for unlawful conduct by bodies other than courts. This can slow down enforcement action and have led to calls for speedier enforcement mechanisms, such as the establishment of the UK Intellectual Property Enterprise Court to provide faster and more cost effective resolution of intellectual property disputes.

In the context of regulating online harms, policy makers have explored ways to impose administrative sanctions on online service providers.

Examples of the tool

Ireland

The Irish Law Reform Commission described the power to impose administrative sanctions as one of the most effective in the regulatory toolkit, and that "the power to impose administrative financial sanctions is both valuable and necessary in ensuring that financial and economic regulators have the requisite powers to achieve their regulatory objectives."

Ireland's online safety proposals are considering the imposition of administrative sanctions by the new Online Safety Regulator, and developing a constitutional approach to enable fines or civil penalties issued to be subject to later endorsement by a court if required.[17]

Tool 15: Executive accountability

What is the tool?

Executive accountability regimes are found in financial services regulatory regimes, including in Australia and the UK. Their objective is to drive culture change in risk management. These regimes sheet responsibility home to key executives if there is organisational conduct that falls short of required standards. This can involve civil penalties such as personal fines and even extend to criminal sanctions including jail time.

[17] See https://www.dccae.gov.ie/en-ie/communications/legislation/Pages/General-Scheme-Online-Safety-Media-Regulation.aspx

Examples of the tool

Australia
The Australian financial services regime has for some time included a banking executive accountability regime (BEAR) that establishes accountability obligations for authorised deposit taking institutions and their senior executives and directors.

Executive accountability was recently extended to senior executives of online providers through the controversial Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act. In addition to heavy penalties for corporations, the penalty for recklessly failing to remove abhorrent violent material as committed by an individual, includes potential imprisonment of up to three years, a fine of up to $2 million, or both.

United Kingdom
The UK is considering implementing executive accountability as a response to major breaches of the statutory duty of care (see below). This could involve personal liability for civil fines or even criminal liability.

It remains to be seen whether the UK will implement executive accountability and if so, which positions within an organisation it would attach to and whether it would apply to companies of all sizes, i.e. including small businesses. The Government is expected to reach a decision in the coming months.

Ireland
Although Ireland is considering the imposition of administrative sanctions as a primary enforcement mechanism, the Minister for Communications has said that holding individuals accountable was still possible under the blueprint for reform:

> *Failure to act on the Online Safety Commissioner's proposals will be a criminal offence ... [and] the concept of individuals being prosecuted under this legislation is still very much open.*[18]

Tool 16: Disruption of business

What is the tool?

There is discussion in some countries of moves to disrupt the business of online providers who are not taking adequate measures to prevent harmful content and conduct online. The objective is to reduce the provider's capability to engage with end users.

In certain circumstances, Governments have considered the ultimate business disruption, blocking an Internet business from the Internet for failing to comply with local laws and online safety regulations.

[18] See https://www.rte.ie/news/technology/2020/0110/1105465-online-safety/

Examples of the tool

United Kingdom
The Online Harms Consultation Paper raised the prospect of forcing "third party companies to withdraw any service they provide that directly or indirectly facilitates access to the services of the first company, such as search results, app stores or links on social media posts." This response would be contained to extremely serious breaches "such as a company failing to take action to stop terrorist use of their services".

In its response to the Paper, the Government has indicated that it is still considering implementing these measures with a decision expected in coming months.

Australia
The Online Safety Act Consultation paper proposes a new 'ancillary service provider notice scheme'. This Scheme would cover service providers that are not directly responsible for the publication of harmful content and conduct online. It would enable the eSafety Commissioner to request (but not require):

- search aggregators to delist or de-rank websites that have been found by the eSafety Commissioner to be "systemically and repeatedly facilitating the posting of cyberbullying or cyber abuse material, image-based abuse or hosting seriously harmful content"; and
- digital distribution platforms to "cease offering apps or games found by the eSafety Commissioner be systemically and repeatedly facilitating the posting of cyberbullying or cyber abuse material, image-based abuse or hosting illegal or harmful content".

This would place expectations upon App stores to remove content identified by the eSafety Commissioner as being seriously harmful or containing cyberbullying/abusive material. There would be no penalties for noncompliance with this notice scheme, but the eSafety Commissioner would be empowered to publish reports on service providers who had failed to respond.

These powers are intended to be used as 'reserve powers' in relation to ancillary service providers where more direct take-down powers used against the primary providers of harmful material have not been effective.

Ireland
The proposed Irish *Online Safety and Media Regulation Bill* will establish a number of codes to address online harms. As discussed in tools 18 and 19, consideration is being given to administrative and criminal sanctions for non compliance with online safety standards. A further sanction being considered is blocking an offending service in Ireland (for example, if an online service repeatedly fails to comply with online safety expectations, the Online Safety Commissioner would have power to compel ISPs to block that service from being able to be accessed by Irish Internet users).

Tool 17: Filtering / ISP level blocking

What is the tool?

Internet Service Provider (ISP) level filtering involves Internet service providers blocking access to certain URLs (website addresses) to ensure that Internet users in the ISP's country can't access the site (subject to using technological workarounds).

There are many complexities in how ISP level filtering might be implemented, including the extent of blocks, technical constraints on its effectiveness and the scope for erroneous inclusion on filtering lists. These complexities are explored in more detail in InternetNZ's paper *To block or not to block - Technical and policy considerations of Internet filtering*.

Examples of the tool

Australia

Australian ISPs have been offering certain limited forms of filtered service for some time. Predominantly this is in relation to child sexual exploitation material, including through ingesting lists of URLs from organisations such as the Internet Watch Foundation. Also, since 2002, certain Australian Internet service providers have offered a family friendly Internet service to those users who select it.

Following the 2019 terrorist attacks in Christchurch, ISPs voluntarily blocked access to sites known to host footage of the attacks and the manifesto of the alleged perpetrator. This was a challenging and problematic approach given the lack of regulatory backing for the action. Following the Christchurch Attacks, the Australian government convened the Taskforce to Combat Terrorist and Extreme Violent Material Online, and one action arising was for the eSafety Commissioner to work with ISPs to put in place a protocol that supports ISPs to block websites hosting graphic material that depicts a terrorist act or violent crime for the period of time directed by the Commissioner. This protocol has now been implemented.

The Online Safety Act consultation proposes to establish a "specific and targeted power for the eSafety Commissioner to direct ISPs to block certain domains containing terrorist or extreme violent material, for time limited periods, in the event of an online crisis event." The proposal would see ISPs provided with civil immunity from any action or other proceeding for damages as a result of implementing the requested blocks. It would also put in place notification and appeal mechanisms.

In the copyright context, Australia has introduced a site blocking regime which allows rights holders to apply to the Federal Court to obtain injunctions to require ISPs to block websites that have the primary purpose or effect of infringing, or facilitating an infringement, of copyright. Australia recently extended this ISP level

site blocking regime to include a 'search filtering' obligation.[19] Rights holders can now apply for an injunction that would both require ISP blocking of infringing websites, and also an order that would require an online search engine provider to take such steps as the Court considers reasonable so as not to provide a search result that refers users to the blocked online location.[20]

United Kingdom

ISPs in the UK work with the Internet Watch Foundation to take action against child sexual exploitation material. The Online Harms White Paper proposed further ISP blocking, however the Government's response to that consultation indicates that they will not progress this initiative. Rather, the voluntary CSEA filtering that currently happens will continue.

The UK consultation had raised the possibility of ISP blocking non-compliant websites or apps, essentially blocking certain services from being accessible in the UK, as an enforcement option of last resort. It would only have been used where a company had committed serious, repeated and egregious violations of the requirements for illegal harms after repeated warnings and notices of improvement.

Israel

Israeli ISPs are required by law to present Internet users with content filtering software solutions to provide Internet users with the means of limiting access to harmful content online. However, research conducted in 2017 by the Knesset Research and Information Centre based on data provided by ISPs showed that only 0.1 - 1.5% of Israeli Internet users made use of content filtering software.

Singapore

ISPs in Singapore are regulated by the Media Development Authority (MDA) which requires service providers to obtain a licence and comply with licence conditions and an Internet Code of Practice. ISPs are required take "all reasonable steps" to filter any content that the regulator deems "undesirable, harmful, or obscene."

The MDA describes this requirement as follows:

> As a symbolic statement of our societal values, local ISPs are required to restrict public access to a limited number of mass impact websites which contain content that the community regards as offensive or harmful to Singapore's racial and religious harmony, or against national interest. The majority of the websites on the list are pornographic in nature.[21]

Tool 18: Duty of care

What is the tool?

Tort law puts a duty of care on certain individuals in certain circumstances to take reasonable care not to cause foreseeable harm to others. In recent times there

[19] Copyright Amendment (Online Infringement Act) 2018 https://www.legislation.gov.au/Details/C2018A00157
[20] Subsection 115A(2) *Copyright Act 1968*
[21] IMDA Internet Regulatory Framework, https://www.imda.gov.sg/regulations-and-licensing-listing/content-standards-and-classification/standards-and-classification/internet

have been explorations of creating a legislative duty of care for online service providers to encourage them to be more proactive in responding to safety risks on their services.

Examples of the tool

United Kingdom
Following its Online Harms White Paper consultation, the UK plans to impose a statutory duty of care which will be overseen and enforced by a regulator, most likely OFCOM. The objective of imposing a statutory duty of care is to make companies take more responsibility for the safety of their users and tackle harm caused by content or activity on their services.

The regulator will set out how companies should discharge the duty of care in a code / codes of practice. The code(s) will include requirements for companies to have clear terms of service setting out what is acceptable, and enforce those terms consistently and transparently. This will include removing illegal content expeditiously and having systems in place that minimise the risk of it appearing on the service. A higher level of protection will be required for children.

It is intended that the duty of care approach will encourage companies to have a robust understanding of the risks associated with their services and to take reasonable and proportionate steps to mitigate the risks. When assessing compliance, the regulator will be expected to consider whether the harm was foreseeable and therefore what the reasonable steps would be in the circumstances in order to discharge the duty of care. In the event of a new risk, a company should notify the regulator and discuss the best approach to mitigate it.

There will be differentiated expectations for illegal content and activities, and for conduct that is not illegal but has the potential to cause harm. As such, companies will not be forced to remove specific pieces of legal content. The regulator will also have a legal duty to have regard to innovation and users' rights online, including privacy and freedom of expression.

France
In France's work to embed civil servants within Facebook, the resulting report posits "creating a duty of care from the social networks towards its members". The report states:

> *"In the financial sector, governments have attempted to promote the credible and long-term commitment by financial institutions to actively contribute to achieving the public interest objectives of combating money laundering, drug trafficking and the financing of terrorism. Banking supervisory authorities have therefore devoted their efforts to imposing and monitoring obligations of means, i.e. compliance with certain preventive rules, rather than punishing failures when the risks being combated materialise (without prejudice to criminal proceedings in that case). Therefore, the banking supervisory authorities do not intervene when it is found that a financial institution has been the channel for channelling funds used for unlawful purposes, but when it finds that a financial institution is not implementing a prescribed prevention measure, regardless of whether or not the financial institution is implicated in unlawful behaviour. This intervention approach is designed to create targeted incentives for platforms*

to participate in achieving a public interest objective without having a direct normative action on the service offered."

Tool 19: Criminalising specific types of online content

What is the tool?

Some Governments have introduced criminal sanctions against specific types of harmful online content, including 'fake news' and disinformation as well as the sharing of non-consensual sexual imagery.

Examples of the tool

Singapore

Singapore recently introduced the Protection from *Online Falsehoods and Manipulation Act 2019* (**OFMA**).[22] The law criminalises (and imposes other civil sanctions and orders) any false online statements of fact. A statement is considered to be false "if it is false or misleading, whether wholly or in part, and whether on its own or in the context in which it appears".[23]

Power is vested in *any* Minister (not just for example the Minister with responsibility for Communications or Justice) to decide whether online content is 'false' and to exercise powers under the Act. Ministers can only act against false statements if it is in the public interest to do so. Ministers have power to order corrections dictated by the Government, and power to compel publication of corrections in a variety of locations, from social media platforms and other online locations, to publication of correctional advertising in newspapers or other print publications at the expense of the person making the correction.[24]

The OFMA grants the power to declare websites, social media platforms and other online locations as "declared online locations" if there are 3 or more false statements made on the location in a 6 month period. It is a criminal offence to earn money from a declared online location (including advertising revenue), and it is a criminal offence for anyone to "financially support, help or promote the communication of false statements of fact" on the page.[25]

In addition to these criminal sanctions, the OFMA uses a combination of many other regulatory tools outlined in this paper, including:

- Order ISP blocking of access in Singapore to the online location that made the false statement[26];

[22] See Singapore Ministry of Law, "FAQ: Protection from Online Falsehoods and Manipulation Act 2019" <mlaw.gov.sg> at June 2020
[23] Subsection 2(2)(b) Protection from Online Falsehoods and Manipulation Act 2019
[24] See Part 3
[25] Part 5 ss 32, 36 and 38.
[26] Part 3 s16(2) and Part 5 s33(3)

- Order online locations (including websites, or social media or communications platforms like Facebook, WhatsApp or Twitter) to notify Singaporean users (in terms dictated by the Government) that a statement on their platform is false[27];
- Order online locations to disable access to false content; and
- Order ISPs to block access in Singapore to online locations that fail to comply with orders regarding false statements[28].

Tool 20: No regulation/positive regulation

What is the tool?

This is not a tool in the same family as the other tools above, but is an approach that could be considered prior to, instead of, or in combination with other tools.

In some instances, Governments have deliberately prioritised a de-regulatory agenda, or Internet regulation has been considered to be contrary to other civil or political values.

In other circumstances, the benefits to society from regulatory options do not outweigh the costs of regulation. For example, the Organisation for Economic Co-Operation and Development (**OECD**) recommends that member countries, when conducting regulatory impact assessments, consider means other than regulation and identify the trade-offs of the different approaches considered. The 'no regulation' option, or baseline scenario, should always be considered.[29]

In these instances, Governments may choose to focus on policies that prioritise growth of the digital economy so as to create a positive digital environment.

Examples of the tool

Estonia

The Ministry of Economic Affairs and Communications adopted the Digital Agenda 2020[30], which focuses on creating an environment that facilitates the use of ICT and the development of smart solutions in Estonia in general. The Government's focus has been on initiatives such as completion of an ultra-fast broadband fibre optic cable network and a 5G activity plan.

Regulatory attention has prioritised:

- Adoption of AI applications in the public sector
- Technological and legal conditions created so people have control of their data in the hands of the state
- Cyber security capabilities strengthened
- E-governance innovation to be accelerated
- E-residency program to be expanded.

[27] Part 4 s21(1)(a)
[28] Part 4 s28(2)
[29] OECD, *Recommendation of the Council on Regulatory Policy and Governance*, March 2014, p26.
[30] See https://e-estonia.com/

United States

Several attempts to regulate the Internet have been struck down as in violation of the First Amendment to the US Constitution regarding freedom of speech. For example, in *Reno v American Council for Civil Liberties*, the Supreme Court held that legislative restrictions on both the "display" and "transmission" of indecent communications online violated the First Amendment.

> "Through the use of chat rooms, any person with a phone line can become a town crier with a voice that resonates farther than it could from any soapbox. Through the use of Web pages, mail exploders and newsgroups, the same individual can become a pamphleteer ... the content of the Internet is as diverse as human thought ... [there is] no basis for qualifying the level of First Amendment scrutiny that should be applied to this medium."

In response to the decision, Congress passed the Child Online Protection Act (COPA), which addressed childrens' access to commercial pornography and described methods to be used by site owners to prevent access by minors. However, COPA was also struck down for First Amendment reasons (Ashcroft v. American Civil Liberties Union), with the Supreme Court arguing that less restrictive methods on speech (such as filtering or blocking technologies) should be used instead. Broadly similar requirements apply in New Zealand, requiring that regulation impacting human rights interests only does so in a way that is legal, necessary, and proportionate.

Country context analysis

Country	Population [31]	Nominal GDP (Q4 2019)[32]	Freedom on the Net ranking[33] 2019 (100 = most free, 0 = least free)	Internet access/ uptake 2018
New Zealand	5.0m	USD$51b	Not included	89% of the population has access to the internet
Australia	25.3m	USD$345b	77	89% of population has access to the Internet
UK	66.4m	USD$718b	77	90% Internet coverage
Ireland (Member of EU)	4.9m	USD$97.4	Not included	89% of population has access to the Internet
Canada	37.6m	USD$483b	87	94% of population has access to the Internet
European Union	446m	USD$3.8 trillion	N/A	89% of population has access to the Internet
Germany (Member of EU)	83.6m	USD$962.5b	80	91% of population has access to the Internet
France (Member of EU)	65.2m	USD$647.1b	76	82% of population has access to the

[31] Taken from official statistics data eg Australian Bureau of Statistics and UK Office for National Statistics

[32] https://www.ceicdata.com/en

[33] Freedom House, Freedom on the Net 2019, an annual ranking of Internet freedom supported by the U.S. State Department's Bureau of Democracy, Human Rights and Labor (DRL), the New York Community Trust, Google, Internet Society, and Verizon Media https://freedomhouse.org/sites/default/files/2019-11/11042019_Report_FH_FOTN_2019_final_Public_Download.pdf

				Internet
Singapore	5.7m	USD$91.7b	56	84% of population has access to the Internet
Estonia (Member of EU)	1.3m	USD$7.8b	94	90% of population has access to the Internet
Israel	8.8m	USD$104.3b	Not included	Just under 80% of population has access to the Internet
United States	331.0m	USD $5,432.3b	77	73% of Americans have home Internet connections[34]
South Korea	51.3	USD$410.9b	64	95.9% of South Koreans have access to the Internet[35]

[34] https://www.pewresearch.org/internet/fact-sheet/internet-broadband/
[35] https://www.statista.com/statistics/255859/internet-penetration-in-south-korea/