To block or not to block

Technical and policy considerations of Internet filtering

internetnz

Table of contents

Introduction

InternetNZ works for an Internet that is open, secure and for all New Zealanders. The underlying technologies that make the Internet function have enabled innovation, personal and economic growth, and access to information. Unfortunately, they have also enabled people to use the Internet to cause serious harm. The idea of using technology to block and filter harmful content may appeal to policy makers looking at options to reduce this harm.

This paper offers an overview of blocking and filtering measures, assessing options in terms of:

- **Technical concerns**: what works, how effective is blocking, and how does it affect the functioning of the Internet?
- **Policy concerns**: what are the impacts on how New Zealanders can use and benefit from the Internet? Are there human rights concerns?

Content by any other name smells just as sweet

One of the challenges around "content blocking" is that it's a very abstracted term. We encourage you to not let yourself think about "content" and instead think about "activity" — the things New Zealanders do online. People post things, share photos, videos, blog, debate and argue, find out information, news, and so on. What is labelled "content" is actually expression, creativity, engaging with family and personal exploration. Using language that abstracts from this activity can make it easy to ignore the real people affected by Internet filtering and blocking. Don't lose sight of how New Zealanders use the Internet, and how important that use can be, both to them as individuals, and to wider society.

"Blocking content" can mean different things to different people

At times in this paper, we refer to blocking content and filtering content. The catch-all phrase in common use is "content blocking," but it is important to note that the actual processes are more complex. Blocking content for New Zealanders will, in most cases, involve routing all Internet traffic through a type of filter, which will scan traffic against a list of predetermined blocked URLs, domain names or keywords. If the filter finds someone trying to reach a page listed on the filter, they may send the user to a 'stop' page, a different website than intended, or some version of a "page not found" message.

When we talk about content blocking, we are using the phrase (for the purposes of this paper) to refer to any combination of the above processes.

A framework for evaluating content blocking

When thinking about Internet filtering and content blocking we need to consider:

- the problem to be addressed
- how the technology works (and whether it can solve the problem)
- the policy impacts (including the harm or unintended consequences that could result from blocking).

Technical issues: can blocking hit targeted content?

Which content is targeted, and how effective blocking can be, will depend on the policy problem. Is the problem driven by people in New Zealand accessing content that is illegal or harmful (eg objectionable material, copyright infringement or online gambling), or by bad actors who expose New Zealanders to harmful content (eg the Christchurch attack livestream, spam, malware, and viruses)? Each policy problem raises technical challenges, for example:

- **Circumvention is often easy** for people who want to access content, which reduces the effectiveness of blocking as a policy measure.
- **Blocking leaves content online**, and does not directly affect the source, so it has limited — if any — impact as a punishment or deterrent.
- **Blocking is hard to target**, as technical blocks are blunt tools which can impact non-targeted content. This may be through blocking entire domain names or IP addresses which also deliver non harmful content, or blocking false-positives, non harmful content which is read as harmful by an algorithm or filter.
- **Blocking methods can create security risks**. Methods that reroute DNS traffic can open the system up for abuse by attackers.

Policy issues: what harms will result from blocking?

Blocking measures may directly cause unjustifiable harms to New Zealanders:

- **Blocking can have collateral damage**, blocking access to content that New Zealanders should be able to access, for example, capturing journalism or research which deals with targeted content.
- **Blocking may require mass surveillance**, harming New Zealanders' privacy rights, for example, deep packet inspection requires reading the content of all web traffic, including emails and login information.
- **Blocking harms New Zealanders' free expression** by limiting people's freedom to find, use, and share information online.

Blocking measures may indirectly cause harms to New Zealanders:

- **Blocking may inspire retaliation**, targeting New Zealand's people, institutions, and computer systems. Blocking is controversial and risks a "whack-a-mole" effect where harmful content proliferates.
- **Blocking may damage New Zealand's reputation**, due to its impacts on human rights and the potential for unintended side effects. As a result, New Zealand may lose international respect and influence.

Governments cannot restrict citizens from accessing content on the Internet without the compliance and cooperation of intermediaries. Internet filtering can happen at different points along the network, whether on a user's device, through an Internet Service Provider, or on a website. It is critical to understand the ramifications of each intervention point.

Where can content restriction happen?

User

Alice creates objectionable content

Bob may not see Alice's content if he has a browser extension that blocks keywords he has chosen

User

Local network

Alice uploads content to a website, through an app or a platform

Bob's workplace is using custom network restrictions (URLs manually blocked by admin)

Bob's family is using an opt-in commercial filter (ie to block children from explicit material)
Network for Learning - is a network NZ schools use which filters content based on the school's needs

Local network

Internet shutdowns

Internet infrastructure

Domain name registry suspends domain name registration of website hosting Alice's content
InternetNZ suspended a domain after Christchurch and unsuspended once objectionable content was removed.[1]

Website/content delivery host **suspends hosting of Alice's website.**[2]
Cloudfare suspended The Daily Stormer

Bob's Internet traffic is diverted through a filter for URL blocking, which can target specific pages or elements within a website, so the core website is still accessible
New Zealand's Digital Child Exploitation Filtering System works like this.[3]

Bob's ISP has implemented keyword based blocking through deep packet inspection
The Chinese firewall works like this - ISPs inspect every bit of data over the network and blocks content with certain keywords in it.[4]

Bob's ISP implements IP based blocking, or DNS poisoning to prevent Bob reaching Alice's website
In 2008, Pakistan blocked YouTube using this method, broadcasting a false IP address that was shared to the world and brought down YouTube for everyone for several hours.[5]

Internet infrastructure

The use of encryption, or a VPN (virtual private network) would circumvent many methods of content restrictions at the local network and infrastructure level.

Social media platform

Social media platform removes Alice's content, through one of the methods below:

Community guidelines: set expectations for users and moderators about what content is allowed on a platform

Upload filters: Image hash database prevents upload of previously banned images

Users report content they deem offensive/harmful

Manual moderation

Social media platform

1 https://dnc.org.nz/sites/default/files/2019-06/201819_transparency_report%20v0.2.pdf
2 https://blog.cloudflare.com/why-we-terminated-daily-stormer/
3 https://www.dia.govt.nz/Censorship-DCEFS-Public-Information-Pack/
4 https://www.newscientist.com/article/mg24132210-400-chinas-great-firewall-and-the-war-to-control-the-internet/
5 https://www.cnet.com/news/how-pakistan-knocked-youtube-offline-and-how-to-make-sure-it-never-happens-again/

Content restriction: reviewing the options

Giving Internet users control: empowering who uses the Internet

A major reason for restricting content is to protect children or prevent users from being unwillingly exposed to harmful content. We can circumvent some harm by enabling Internet users to control their own experience online through education and tool provision.

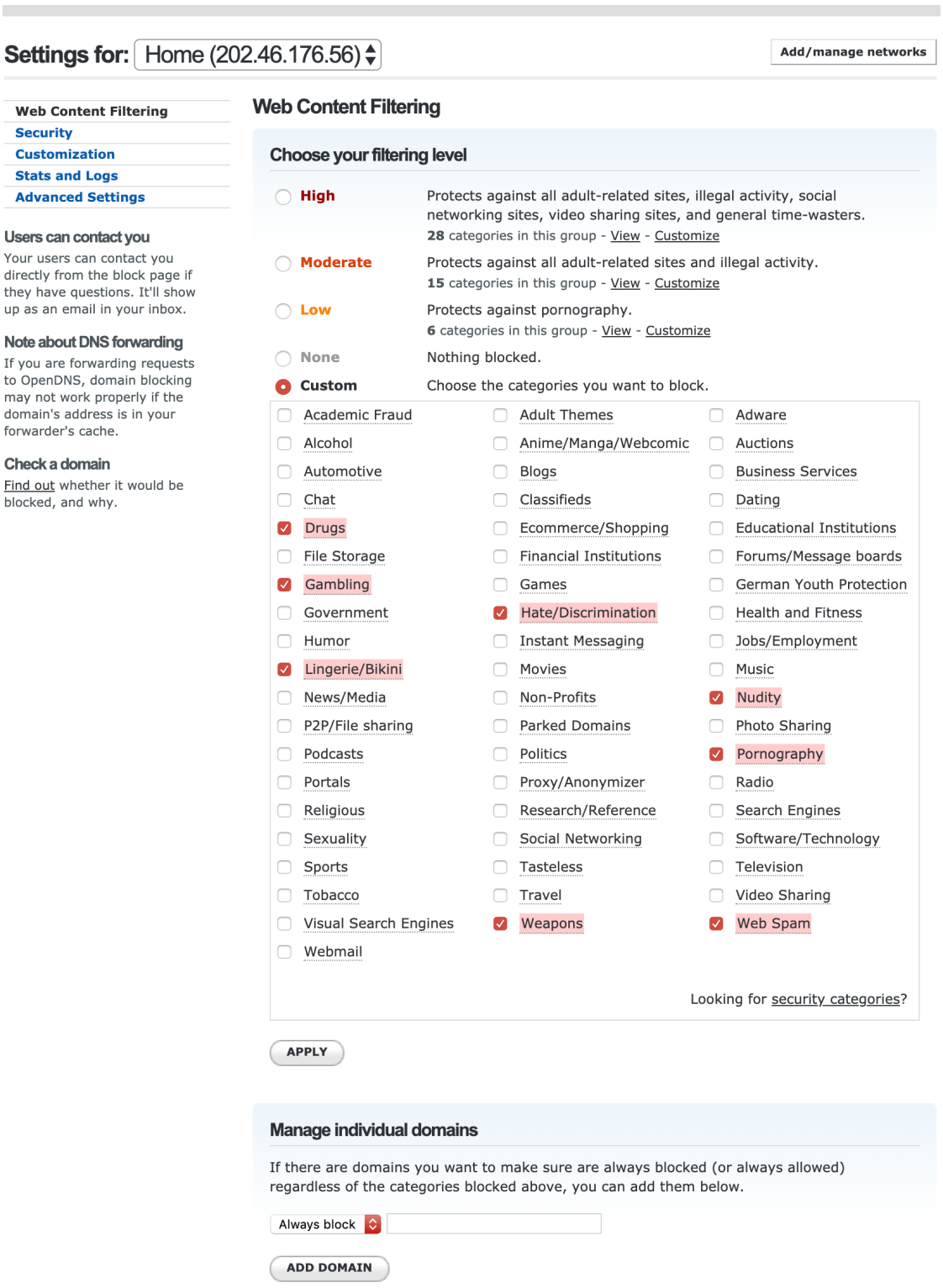
Opt-in tools give users control, allow transparency about what is being blocked, and empower users to think critically about the experience they want for themselves and their children. These tools can be implemented at the network level, so that any device connecting to your home network will be affected, or at the device level, so anyone using a specific device would be affected.

This is the most effective option if you are trying to stop people inadvertently seeing harmful content. This option is likely to have limited effectiveness against people intentionally seeking out harmful content.

Commercial home Internet filters

Set on a device, or a home router, these opt-in filters are commercial products that broadband customers can implement. These products use lists of pre-categorised web addresses to filter out unwanted content.

Figure 1: example home filtering tool



Preconfigured DNS services

Recursive DNS services like OpenDNS can be used instead of your Internet Service Provider's default settings. They reroute traffic through a filter that can be configured by the user. Domain names are tagged by category, and then the administrator can choose which categories are blocked for the household.

- This filtering is effective in stopping children from reaching harmful websites but is ineffective in blocking harmful content on media aggregators, like Google image search and all social media platforms. You can only block all of google.com, not just parts of it.
- It is ineffective in filtering harmful content being delivered in apps including individual posts on feed based apps like Twitter, Facebook or Tumblr.

Browser extensions

A person's experience on the Internet is infinitely customisable. If an Internet user wants to access content, they have many avenues to do so. If a user wants to proactively protect themselves from content online, whether it is objectionable, abusive, or just a TV spoiler, they have many options to do this. Examples include:

- Share No Evil (see the next page for more information)
- ad blockers (many people block ads either for usability or for cybersecurity reasons)
- anti-distraction extensions such as StayFocusd which block you from social media and shopping sites when you are focussed on a particular task.

Share No Evil: Shielding yourself from specific harms

Share No Evil is a browser extension released in 2019 in response to the March 15 terror attack on Christchurch mosques.[6] The extension scans a page for instances of the name of the alleged terrorist, and block the name from the user. While this is largely a symbolic gesture, it is an example of a user controlling their online experience through customisable filtering at the browser level.

6 Colenso BBDO (2019), 'Share No Evil', https://sharenoevil.co.nz/

We recommend options that give people the information and control to make the best decisions for them and their families. This can allow people to avoid some harmful content without compromising the integrity of the Internet.

Recommendations:

- **Educate New Zealanders about how they can use home filters effectively, to make filtering choices for themselves and their families.**
- **Encourage use of opt-in filters and tools.**

Social media platforms: moderating what happens on the Internet

Much of the content that governments may want to restrict is user-generated content hosted and shared widely on platforms. Websites and services like Facebook, Google search, YouTube, and Wikipedia enable people to share content globally in ways that can reach billions of people. These platforms have levers for boosting the reach of content or restricting access to it in ways that can be more effective than blunt tools like IP or DNS block lists.

Content shared on social media platforms can cause immeasurable harm, and the platforms themselves can be utilised for mass diffusion of harmful content to a wide audience, but blocking access to these platforms as a whole is not an acceptable remedy.

Social media and platform moderation is the ideal place for making content decisions when the consideration is integrity of core Internet infrastructure.[7] However, it needs to be done in a way that is transparent, accountable, and allows for due process.

As with society in general, there will always be fringe websites and platforms online that will have niche audiences and may share content many New Zealanders deem inappropriate or harmful. In some cases, blocking content may not be as detrimental to New Zealanders' Internet experience, but it will also do little to reduce exposure to harmful content.

We think controlling sharing and reach through popular online platforms is a possible and effective way to mitigate their impact.

7 All hardware and software systems that constitute essential components in the operation of the Internet. See more at https://www.cyberpeace.org/critical-internet-infrastructure/

Community guidelines

Hosts of user-generated content (UGC) often have terms of service and community guidelines they seek to uphold. These guidelines are often informed by the creators' biases and imbue the values of the culture they are created in. Guidelines are not always straightforward. Take for example the implementation of guidelines around breastfeeding images.

Facebook and Instagram's guidelines exclude images of female-identified nipples but allow images of breastfeeding, art, or historical significance. This has lead to many enforcement difficulties.[8] For example:

- **False positives**: The AI often flags images that are not of nipples, or removes images of breastfeeding.
- **Culture gap**: Moderation guidelines are written from a specific cultural context. What is acceptable in the United States may not be acceptable in New Zealand (or vice versa), yet we are subject to the guidelines written for US cultural norms.

Community guidelines are easy to write, but difficult to implement, enforce, and gain the support of a platform's users. The two main ways that community guidelines are enforced are through automated and human moderation.

Automated upload filters

With an upload filter in place, any content uploaded to a platform will be checked against a database of keywords and image hash (a process of creating a unique key associated with an image), to assess whether the content could be harmful.

The concerns about upload filters:

- **Surveillance** — Upload filters require all content to be inspected and can be considered a kind of constant surveillance.
- **Oversight** — There is a lack of transparency in how these databases work, and their effectiveness, proportionality and appropriateness to achieve the goals the platforms aim to achieve.
- **Ineffectiveness** — Filters are unable to understand the context and are limited to what is already in the database, so they are easy to evade in some instances and cause too many false positives in others.

8 The Verge, 'Instagram will now warn users close to having their account banned' 18.07.2019, https://www.theverge.com/2019/7/18/20699393/instagram-account-ban-warning-message-moderation-update

GIFCT Hash Database

The Global Internet Forum to Counter Terrorism is a coalition of tech platforms who work together to reduce the spread of terrorist content on platforms. They have a shared industry hash database which now contains more than 200,000 hashes, where companies can create "digital fingerprints" for terrorist content, remove matching content and, in some cases, block terrorist content before it is even posted.[9]

The Forum is limited to platform companies, and only the founding members get voting rights. There is no transparency as to how content is added to the database, or how many false positives are removed from platforms.

A hash database is easy to circumvent. Videos and images can be manipulated by (for example) adding a watermark, speeding up the video, or adding a border, all of which render the hash useless.

9 York, Jillian C., 'Caught in the Net: The Impact of "Extremist" Speech Regulations on Human Rights Content' 30.05.2019, https://www.eff.org/wp/caught-net-impact-extremist-speech-regulations-human-rights-content

Manual content moderation

Many large online platforms rely on human moderation as well as automated moderation. Content might be flagged by a user, and then reviewed by a moderator, who has a few seconds to decide if content should remain on the platform, or be removed. Outsourcing moderation to call centre-like offices has many pitfalls:

- **Localisation** — Moderators can be based anywhere in the world, and may not accurately identify hate speech or abuse when local or niche knowledge is required.
- **Right to appeal** — When users have their content removed, or their accounts suspended, the avenues to appeal decisions are difficult to access and slow to respond.
- **Emotional and psychological toll** — Moderation is low paid, entry level work treated as low skilled. Moderators are subject to long hours of watching, reading and looking at content deemed unsafe and unacceptable for Internet users to see. There is a growing understanding of the toll on moderator wellbeing in this line of work, and it is not sustainable.[10]

Traditional Internet forums, like Reddit,[11] or smaller message boards, often have volunteer moderators who are engaged members and recognised leaders of that community, and they work to enforce the community guidelines and make the forum a place where their members want to be. This method of moderation can create strong social cohesion within the group, and there is buy-in from users who support their moderators. There are some limitations to this approach:

- **Objectivity** — A moderator may be biased against people or ideas, and silence dissenting views. It is hard to hold a moderator accountable if they are acting against the interests of the community they represent.
- **Inability to scale —** A volunteer moderator may be able to moderate a forum of several hundred people, but as membership grows, so does the complexity and size of the job.

10 Newton, Casey, 'The Trauma Floor' 25.02.2019, https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona
11 Reddit, 'Moderation wiki', https://www.reddit.com/wiki/moderation

Blocking through Internet infrastructure: breaking what the Internet is

Content or site blocking can also happen at the infrastructure level where core Internet protocols make connections and deliver information. For our purposes, infrastructure includes Internet Service Providers (ISPs), the Internet Protocol (IP), the Domain Name System (DNS), Content Delivery Networks (CDNs) and the servers that host copies of content.[12]

People connecting to the Internet do so through an ISP which provides their mobile or fixed-line connection. It puts ISPs in the technical position to modify connections to domain names, URLs, and IP addresses. Doing so is technically possible, but violates the end-to-end principle, which requires that connections are controlled by the people using them.[13]

In a country like China, the government can implement blocking through its more direct control of Internet connections. In a democratic nation like New Zealand, where Internet services are provided by independent competitive market players, cooperation from commercial and independent ISPs is needed to implement any of the blocking methods set out on the next page.

12 For a detailed look at how the Internet works, see InternetNZ's Internet Openness: what it is and why it matters (2019). Also see ISOC, 'Content Blocking Overview' (2017), https://www.internetsociety.org/wp-content/uploads/2017/03/ContentBlockingOverview.pdf

13 Jerome H Saltzer, DP Reed and David D Clark "End-To-End Arguments in System Design" (1984) 2 ACM Trans Comput Syst 277.

Infrastructure blocking methods

	DNS filter	IP filter	URL filter	Deep packet inspection
Overview	An ISP may divert Internet traffic through a DNS server which blocks look ups of certain domain names	A filter in the network can block certain IP addresses from successfully returning a response	A URL filter can block certain URLs, which means it can block specific pages within a website without blocking access to the whole site	Blocking based on deep packet inspection is where individual packets of data are inspected to identify whether they contain keywords or components of images, and flagged web pages or elements are blocked from the end user
Dangers of this method	Blocks access to all content served by a domain name, potentially blocking access to non harmful, legal information	An IP filter blocks all content from one IP address, blocking legal, non harmful content as well as illegal content from that address	URL filtering can cause performance problems, decreasing overall speed and reliability	This kind of 'content aware' filtering can cause performance issues for the network as it requires all packets to be passed through inspection engines, introducing network delays
	Can compromise DNS security as the process requires creating vulnerabilities in encrypted traffic	A single IP address can serve many websites (often when a website is hosted on a CMS like Squarespace or Wix) so this method is a blunt tool that can cause collateral damage		False positives are common, as keywords may be acceptable in one context and not in another
Limitations of this method	Users can easily circumvent this method by configuring their routers to using an alternative DNS service (see "Preconfigured DNS server," above)	IP addresses are easy to change, so can be easily evaded by the content publisher IP block lists are often long and hard to maintain due to the ease of evasion	Encryption or use of a VPN renders this technique ineffective	Encryption or use of a VPN renders this technique ineffective Mass surveillance of end users is required
Any instances where this blocking method is justifiable?	ISPs may use DNS filters to protect against security threats such as malware or known phishing domains. These filters are commonplace and non-controversial		This method is the technique used by New Zealand's Digital Child Exploitation Filtering System[14]	

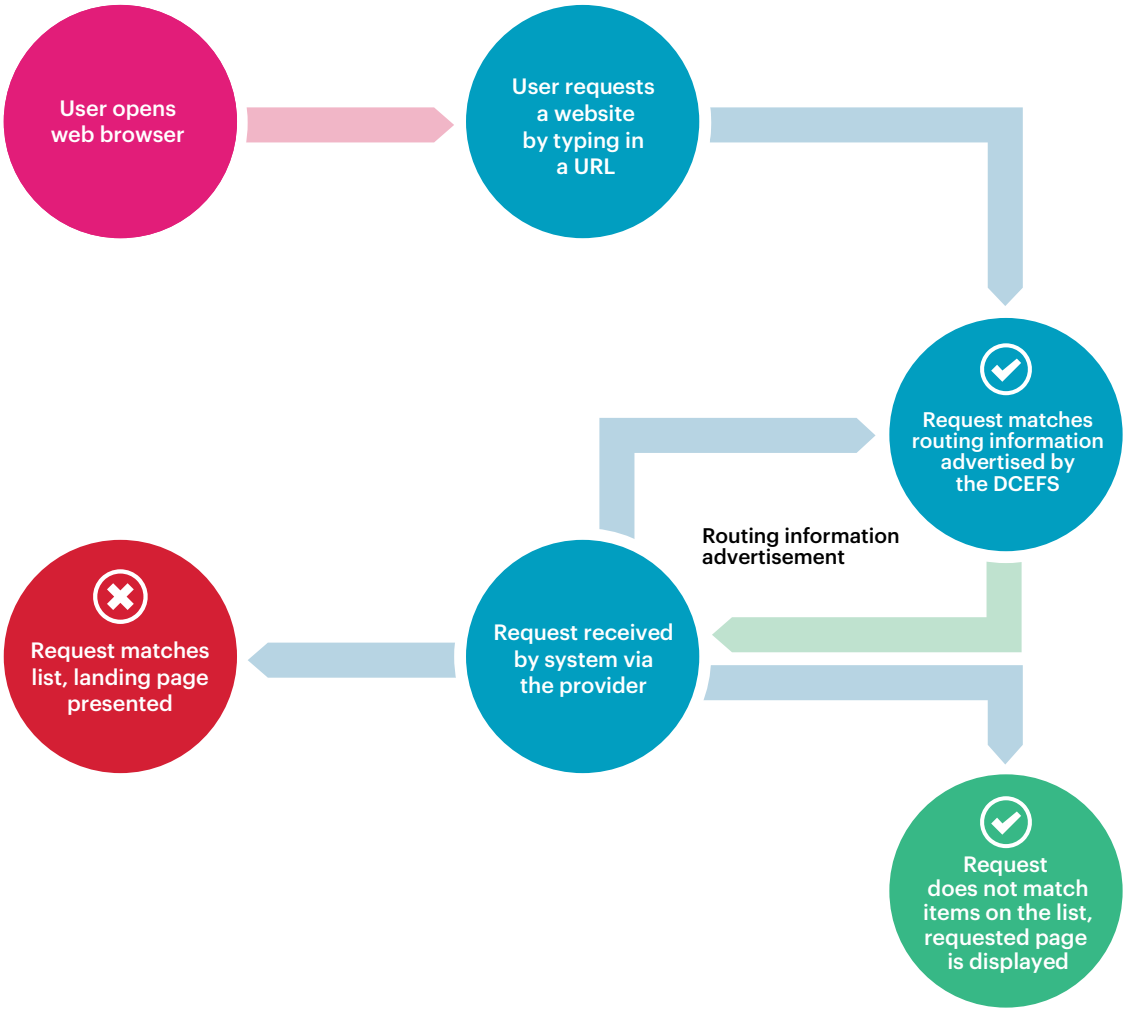
14 Department of Internal Affairs, 'Censorship DCEFS Public Information Pack', https://www.dia.govt.nz/Censorship-DCEFS-Public-Information-Pack#3.

The Digital Child Exploitation Filtering System : Accountability, transparency, and a right of appeal

Since 2010, the New Zealand Government has maintained the Digital Child Exploitation Filtering System (DCEFS).[15] This is an opt-in filter that ISPs can use, which will block a list of websites and pages that contain objectionable material involving children.

The DCEFS is overseen by an Independent Reference Group, a collection of relevant parties who maintain oversight of the operation of the Digital Child Exploitation Filtering System to ensure it is operated with integrity and adheres to the principles set down in the Code of Practice.

15 Department of Internal Affairs, 'Censorship DCEFS Public Information Pack', https://www.dia.govt.nz/Censorship-DCEFS-Public-Information-Pack#3.



Domain name system, website hosts and content delivery networks

In addition to ways ISPs can block content, there are other actors that make the Internet function. Many of these actors provide services, and their users are required to respect the providers' terms of use. In exceptional circumstances, these service providers could assist in restricting access to harmful content, but this should be done transparently and in a way that holds decision makers to account. These providers include:

- domain name registries, like InternetNZ, who operate the top level domain names (.nz), or domain name registrars, who register domains for people to use
- cloud service providers, like Amazon Web Services or Cloudflare.

These services are provided by private companies, which may have terms of service that users must agree to, that go beyond what is required by law.

Cloudflare terminates services for The Daily Stormer and 8Chan

Cloudflare is a business that provides content delivery network services, helping client websites to remain accessible despite outages and attacks, manage traffic and security, and deliver content faster. A key use of their service is to protect against denial of service attacks by third parties which would otherwise make a website inaccessible.[16] In 2017, Cloudflare decided to stop doing business with the alt-right website The Daily Stormer. The CEO was troubled by his ability to make a decision like this, as he recognised the dangerous precedent he was setting:

"the concept of Due Process is close to universal. At its most basic, Due Process means that you should be able to know the rules a system will follow if you participate in that system. [...] Due Process requires that decisions be public and not arbitrary."[17]

According to the Cloudflare CEO, law enforcement, legislators, and courts have the political legitimacy and predictability to make decisions on what content should be restricted. Companies should not.

In August 2019, in response to a mass shooting in El Paso, Texas, Cloudflare terminated its services for 8Chan, the imageboard used by several violent extremists, including the shooter in the March 2019 terror attack in Christchurch.[18]

16 Cloudflare, 'What Is a Distributed Denial-of-Service (DDoS) Attack?' https://www.cloudflare.com/learning/ddos/what-is-a-ddos-attack/
17 Cloudflare, 'Why We Terminated Daily Stormer' 16.08.2017, https://blog.cloudflare.com/why-we-terminated-daily-stormer/
18 Cloudflare, 'Terminating Service for 8Chan' 05.08.2019, https://blog.cloudflare.com/terminating-service-for-8chan/

Human rights and core Internet infrastructure

In the Internet era, using tools that impact the functioning of core Internet infrastructure has important implications for human rights. This includes the right to free expression, affirmed in New Zealand law as the freedom to seek, receive, and share any kind of information in any form.[19] Government actions affecting human rights need to be assessed against a high threshold of:

- necessity
- proportionality
- transparency
- accountability
- due Process from a competent authority.[20]

Government-mandated content restriction by ISPs, when it does not meet the above criteria, undermines New Zealanders' trust in the Internet and the Government.

We think Internet openness depends on ISPs delivering to users the Internet packets that make the Internet work without interference, discrimination or surveillance.[21]

Recommendations:

- **Apply a human rights impact assessment.**
- **Consider less intrusive measures before interfering with Internet infrastructure.**
- **Measures that may impact Internet infrastructure in any way should be time limited and specific.**

19 New Zealand Bill of Rights Act 1990, s 14, http://www.legislation.govt.nz/act/public/1990/0109/latest/DLM224792.html
20 See for example the Ministerial Policy Statement on warrantless surveillance https://www.nzic.govt.nz/. Also see Human Rights Commission, "Privacy, Data and Technology: Human Rights Challenges in the Digital Age" (May 2018), https://www.hrc.co.nz/
21 See for example Jeremy West "A Framework for Understanding Internet Openness" [2016].

Recommendations

Define your policy problem:

Any legislative options need to be designed on a clear problem definition, and the **least intrusive** option that responds to the identified policy problem.

When considering policy options for protecting New Zealanders from harmful and illegal content, we recommend that policy makers think about the following:

- Any action, including steps to block content, must link back to a clearly defined policy problem.
- Are you trying to prevent users from being exposed to harmful content against their will? Or are you preventing users who are looking for this content from finding it?
- Is the host non-compliant to requests for take down?
- Is the harm being prevented worth the risk that the blocking method may cause to the Internet? Make the trade offs very clear — what is to be gained by implementing filters, and blocking content, versus the risk to core infrastructure.

Be clear about proposed technical implementations:

- Any proposal for content filtering or blocking needs to grapple with technical details of implementation, circumvention, and side effects.
- Is the target community likely to circumvent the block i.e. how technically savvy is the target community?

Undertake human rights impact assessments:

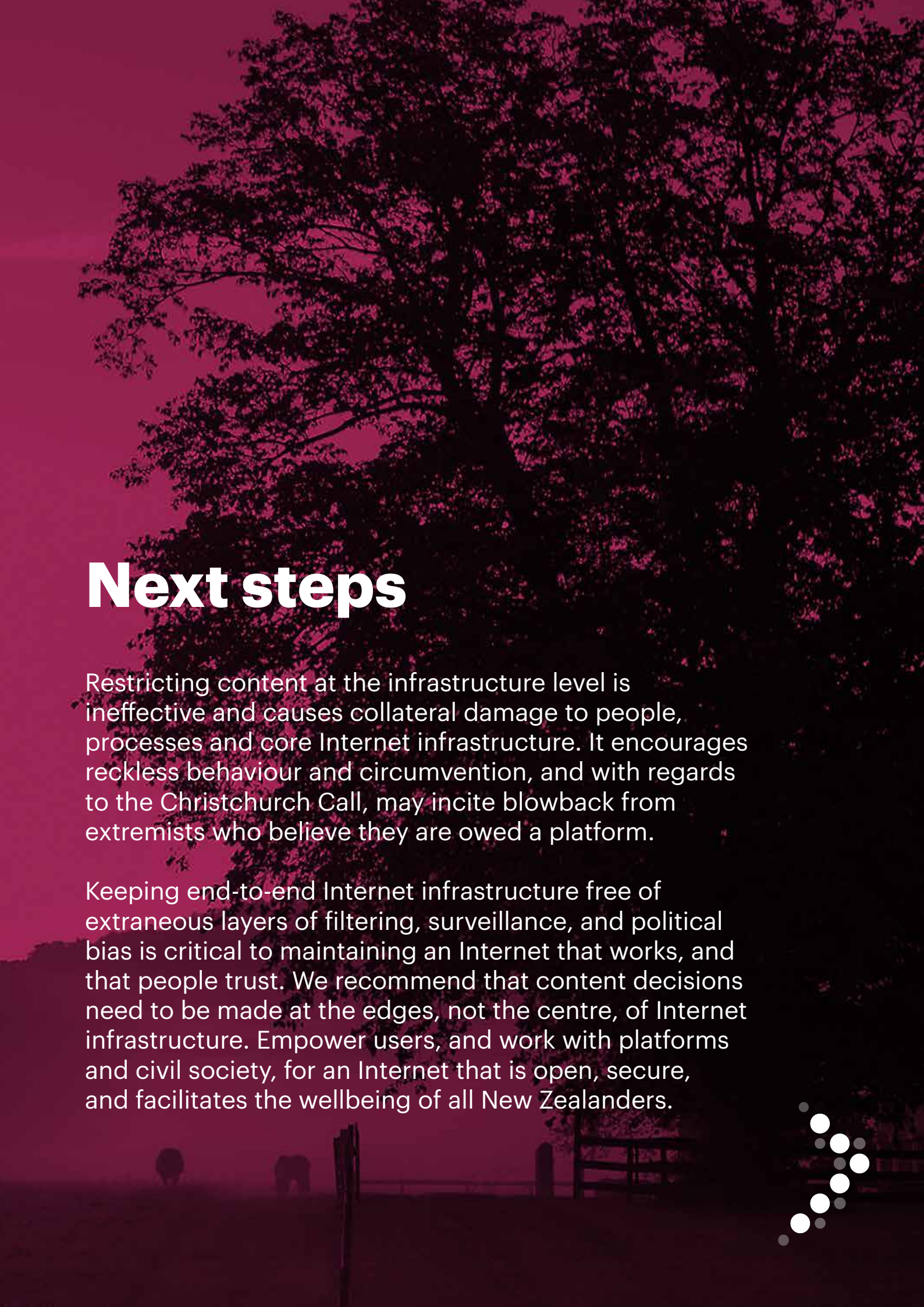
- Review any content policy against a framework of legality, necessity, and
  proportionality.

Encourage users to take control:

- Invest in public education and media literacy.
- Promote personal and local network filters.

Work with online services and civil society:

- Call for platforms to make target content less visible (search and social de-ranking, potentially removal).
- Use open multi stakeholder processes to develop community guidelines that can work for different cultural contexts.
- Encourage content removal at the source. The identification and removal of harmful content is a technical challenge that platforms must strive to achieve.



Next steps

Restricting content at the infrastructure level is ineffective and causes collateral damage to people, processes and core Internet infrastructure. It encourages reckless behaviour and circumvention, and with regards to the Christchurch Call, may incite blowback from extremists who believe they are owed a platform.

Keeping end-to-end Internet infrastructure free of extraneous layers of filtering, surveillance, and political bias is critical to maintaining an Internet that works, and that people trust. We recommend that content decisions need to be made at the edges, not the centre, of Internet infrastructure. Empower users, and work with platforms and civil society, for an Internet that is open, secure, and facilitates the wellbeing of all New Zealanders.

Further reading

Internet filtering and human rights:

- Article 19 'Access to information', https://www.article19.org/issue/access-to-information/
- EFF, 'Caught in the Net: the impact of "extremist" speech regulations on human rights content' https://www.eff.org/files/2019/06/03/extremist_speech_regulations_and_human_rights_content_-_eff_syrian_archive_witness.pdf

More technical information on how Internet filtering works:

- IETF, RFC 7754 'Technical Considerations for Internet Service Blocking and Filtering', https://tools.ietf.org/html/rfc7754
- ISOC, 'Content blocking overview', https://www.internetsociety.org/wp-content/uploads/2017/03/ContentBlockingOverview.pdf

Notes

Notes

Level 11
80 Boulcott Street
Wellington 6011

P.O. Box 11-881
Manners Street
Wellington 6142
New Zealand

Phone: +64 4 555 0123

internetnz.nz

@InternetNZ

InternetNZ

Published September 2019