University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Cohney, S;Scheffler, S;Xiao, M;Habtu, L

Title:
Submission to the Education and Workforce Committee: Inquiry into the harm young New Zealanders encounter online, and the roles that Government, business, and society should play in addressing those harms

Date:
2025

Citation:
Cohney, S., Scheffler, S., Xiao, M. & Habtu, L. (2025). Submission to the Education and Workforce Committee: Inquiry into the harm young New Zealanders encounter online, and the roles that Government, business, and society should play in addressing those harms. New Zealand Parliament, Education and Workforce Committee.

Persistent Link:
https://hdl.handle.net/11343/368849

License:
CC BY-ND

30 Jul 2025

Committee Secretariat
Education and Workforce Committee
Parliament Buildings
Wellington

RE: Inquiry into the harm young New Zealanders encounter online, and the roles that Government, business, and society should play in addressing those harms

Submission from researchers affiliated with The University of Melbourne and Carnegie Mellon University.

Dear Committee,

Thank you for the opportunity to comment on the "Inquiry into the harm young New Zealanders encounter online, and the roles that Government, business, and society should play in addressing those harms".

We are a group of computer scientists writing in our individual capacities as researchers in online security and privacy. Institutional names are for affiliation only.

Young people unquestionably face a wide spectrum of online harms—including privacy invasions, discomfort when encountering adult-themed content, as well as grooming and predation. These encounters vary in type, severity, prevalence, affected age group, possible interventions, and many other factors.

When these disparate harms are lumped together under the banner of "child safety online," policy responses tend to be blunt, poorly balanced, and blind to specific technological limitations and possibilities with narrow implications.

We distinguish harms arising from public content and privately shared content[1] because each demands different technical and policy responses.

Private messages and other privately shared content warrant special treatment. Users—minors included—rightly expect privacy in direct messages. That privacy can shield abuse, but it also enables legitimate expression that would not occur in public.

[1] Note that some harms (consider cyberbullying and non-consensual image abuse) can take place either in public or private, even while they may predominantly take one form.

Furthermore, interventions should be developmentally calibrated: restricting direct messaging functionality for under-12-year-olds is a very different proposition from restricting 14-year-olds.

Interventions

Website Age Restrictions

Many websites already restrict users below threshold ages from accessing the platform. However, enforcement mechanisms vary significantly, with many sites using a perfunctory dialog prompting users to enter their birthdate to gain access. This likely has a minimal deterrent effect. Larger sites, such as major social media platforms, add to this by building mathematical models that use inputs such as someone's posts to additionally detect who may have misrepresented their age. Such a user may then be required to provide government ID as proof of age.

Australia's policy response to this situation is notable: it set a uniform age of 16 as a threshold below which platforms may not host a user. It imposes affirmative obligations on platforms to "take reasonable steps to prevent age-restricted users having accounts".[2] The eSafety Commissioner is tasked with developing guidance as to what constitutes reasonable.

The UK has similarly rolled out the *Online Safety Act 2023*, which, among other requirements, mandates that adult websites use "highly effective age assurance" methods rather than something akin to a cursory birthday check.

Proposed *reasonable* methods (across jurisdictions) include:

- The direct use of government ID; physical or digital
- biometrics, such as face scanning
- sophisticated mathematics (zero-knowledge-proofs) to allow a system that has previously verified a users' age to provably disclose that fact without compromising other user details
- ID-based third-party providers that function without privacy guarantees, or with privacy guarantees only against some parties.

Ofcom endorses several options[3], but many of the methods they endorse have not been empirically proven effective and come with significant costs. The supporting evidence for the accuracy of face-based and other biometric systems[4] is thin, whereas the practical concerns surrounding age-gating are increasingly well documented.

---

[2] Online Safety Amendment (Social Media Minimum Age) Bill 2024 https://parlinfo.aph.gov.au/parlInfo/search/display/display.w3p;db=LEGISLATION;id=legislation%2Fbills%2Fr7284_aspassed%2F0001;query=Id%3A%22legislation%2Fbills%2Fr7284_aspassed%2F0000%22

[3] https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/age-assurance

[4] Existing evaluations have largely been controlled by vendors or other industry insiders with vested interests in the technologies success. Outside evaluations by researchers and government bodies (such as

This is an area the authors of this brief are actively researching, with anticipated results in late 2025/early 2026.

Concerns with Age Gating and Assurance Mechanisms

Age-gating raises numerous problems–both practical and ethical. Many of these are detailed in submissions on the Australian Act[5], including my own team's[6]. Key concerns include:

Speech Chilling Impacts

The imposition of potential liability from allowing underage users to access content/platforms leads to platforms restricting access beyond the facial intent of the legislation–as a method of risk mitigation. This has already led to innocuous sites choosing to shut down rather than deal with the risks and costs of compliance[7] and other sites to block socially important content, such as LGBT+ related health information and discussion, for fear of liability[8]. Promises from the government of the day to only enforce such laws 'reasonably' are not effective in mitigating this as entities reasonably fear how laws on the books can be later weaponized. Such policy implementations have had a demonstrably chilling effect on *non-controversial* adult speech.

Requiring users to upload their ID or scan their face may also chill the speech of legitimate users. Overage users who wish to access speech but are concerned about identity theft or privacy violations may now avoid these websites even though they are legally permitted to access them.

Privacy Implications

There are a variety of privacy harms that may arise from overzealous attempts to solve online safety. These range from an additional party becoming aware of what sites users are attempting to access or patterns of access, to additional storage, processing, sharing, or sale of credential and biometric data–a particular issue for third party age assurance providers.

NIST's FATE https://pages.nist.gov/frvt/html/frvt_age_estimation.html) are yet to legitimate the majority of the OFCOM listed technologies are sufficient robust to a level where it is appropriate to rely on them.

[5]https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Environment_and_Communications/SocialMediaMinimumAge/Submissions

[6] https://www.aph.gov.au/DocumentStore.ashx?id=3f3d60bb-3c53-4757-a81c-2f9a0aa348eb&subId=773513

[7] These include such innocuous websites such as London Fixed Gear & Single-Speed cycling forum (https://www.lfgss.com/conversations/401475/), the Hamster pet forum (https://www.thehamsterforum.com/threads/big-sad-forum-news-online-safety-act.2091/) and a large variety of other websites (see https://www.blocked.org.uk/osa-blocks)

[8] As a result of the UK's Online Safety Act, users in the UK are reporting they can no longer access LGBT+ related communities on reddit unless logged in and verified. https://www.reddit.com/r/transgenderUK/comments/1m8kb4m/reddit_now_censors_lgbtq_content_under_ofcoms_new/

Even with honest providers, large-scale data collection raises baseline risk and invites exploitation by bad actors. The collection of additional data raises risks of more severe data breaches, data trafficking, and accidental disclosures of private information[9].

Further, gating increasing amounts of content behind services that validate identity (or see ID cards or faces) create a baseline privacy threat: they severely diminish the ability for any individual, child or adult, to browse the web privately. In the week since major provisions of the UK act came into force, there was significant public backlash[10].

These identity-based approaches also expose users to greater risk of identity theft, both through the risk of data breach or mishandling by the age assurance vendor itself, and also through desensitizing users to the act of uploading private identification.

Unreliability of Age Estimation Services

Those endorsing biometric/facial age estimation services often claim that their products have high accuracy, precision, and fairness. Under best case conditions (e.g., ideal lighting and framing), these systems still exhibit multi-year errors in guessing someone's age, which rapidly degrades in certain demographics (with average errors of 3+ years in individuals with different skin tones).[11] Further, none of the public evaluations of these systems are in adversarial settings that consider the effectiveness of the systems and algorithms for users who are *intentionally* trying to subvert them, or who may only be able to use these systems under imperfect conditions.

Our research team is currently studying these systems to perform an independent and rigorous evaluation in realistic adversarial settings–e.g., in ways that users will likely use them in practice.

Triviality of Bypassing Geographic Restrictions

Where small jurisdictions impose restrictions on platforms that are not applicable globally, it becomes trivial for users to bypass such requirements using internet technologies that facilitate *tunnelling*. Protocols and applications that support tunnelling, such as VPNs, are easily accessible, cannot be blocked without severe technological consequences, and allow users to access websites while appearing to originate from a different country.

After provisions of the UK Online Safety Act took effect in July 2025, one VPN provider reported a 1,800% spike in daily sign-ups.[12] The pattern of vastly increased use is consistent across the industry[13] which is indicative of both demand and the effectiveness of VPNs for circumvention.

---

[9] We refer the inquiry to reporting and analyses in Wired Magazine, https://www.wired.com/story/the-age-checked-internet-has-arrived/

[10] https://petition.parliament.uk/petitions/722903 with approximately 300k signatures at time of writing.

[11] https://pages.nist.gov/frvt/reports/aev/fate_aev_report.pdf

[12] https://www.ft.com/content/356674b0-9f1d-4f95-b1d5-f27570379a9b

[13] https://www.wired.com/story/vpn-use-spike-age-verification-laws-uk/

Researchers from the University of New South Wales noted (citing a Family Zone study) that "nearly two-thirds of 13- to 15-year-olds"[14] in the study were using VPNs to access gated content. We would expect this number to increase with the presence of more onerous restrictions.

There *are* limited techniques for sites to obtain some level of insight into whether a user is using a tunnel, but this generally does not reveal a user's true location–which prevents a platform from applying that jurisdiction's rules. As a result, it is unlikely that requiring platforms to detect tunnelling will result in the desired policy outcomes. Further, such a requirement would further degrade user privacy and remove one more tool that users can employ to avoid tracking–a common use of tunnels.

Challenges with Setting a Uniform Age

Setting 16 as the minimum age for access to certain services introduces a discontinuity between similarly situated peers, raising both practical and normative challenges. Currently, many platforms set their minimum age for access at 12– when children are assumed to be less mature and less embedded in social-media-centric environments. They are also less likely to be sophisticated as to how they bypass restrictions. By 14, young people are considerably more sophisticated, have greater social needs (which for some may include accessing social support through social media communities), and are also far more capable and likely to engage in circumvention by using technologies such as VPNs.

Addressing "Don't Let the Perfect Be The Enemy of the Good"

Proponents of age-assurance argue that imperfect systems are still worthwhile[15], but this misses the critical point: the outcome of a cost-benefit analysis depends on the actual effectiveness of the technology. A low effectiveness technology/policy with moderate social and financial costs is far less justifiable than a high-effectiveness technology/policy with minimal costs.

It is true that most of the aforementioned options would reduce online harms by making it more difficult for underage users to access online platforms–however, we argue that this would come at a variety of underappreciated costs.

---

[14] See https://www.aph.gov.au/About_Parliament/House_of_Representatives/About_the_House_News/Media_Releases/Parliamentary_Committee_recommends_mandatory_age_verification_for_online_pornography_and_wagering §2.36

[15] See https://www.aph.gov.au/Parliamentary_Business/Committees/House/Social_Policy_and_Legal_Affairs/Onlineageverification/Report §2.39 – 2.41

Antitrust/Competition Concerns

Dominant players in the market may use age-assurance schemes to ensure vendor lock-in and diminish competition. Already, Google has indicated[16] that proprietary components from its Android operating system will be required to use age-assurance features, potentially locking out competitors who build on open-source components. Policy makers should be aware of this dynamic and ensure safeguards are in place to avoid entrenching incumbents.

Age assurance systems can also sometimes prevent classes of overage people from accessing content, for example if their credentials are from jurisdictions a given age assurance vendor does not work with.

Best-in-Class Age Assurance

Device-level restrictions are the *most* effective approach: single-user devices are increasingly shipping with robust, manufacturer-supplied parental controls–age assurance can similarly be implemented across a device or an account on a device.

As most of the relevant individuals now access the internet through single-user devices with improving parental controls[17], restricting access at a device level is not only effective, but minimizes the risk of impinging on others' civil liberties and collateral effects.

It is also challenging for many young people to procure and consistently access a device without a parent's or guardian's knowledge—helping to ensure that, parental controls, where available and configured—limit unfettered access to content.

In practice, device–based restrictions involve integration between the device and operating system, app-stores, and browser restrictions. While this is technically involved, Apple[18] and Google[19] are rolling out technologies to do so. Notably, this can be done in a privacy preserving fashion using a mathematical technique known as *zero-knowledge-proofs*, which became viable for this use case only after recent research. We also direct the inquiry to the report of the New America Foundation on the use of zero-knowledge-proofs for age verification and endorse their recommendations.[20]

---

[16] See https://www.reddit.com/r/degoogle/comments/1mau7yl/eu_age_verification_app_to_ban_any_android_system/ for an indication of the issue

[17] https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/media-literacy-research/children/children-media-use-and-attitudes-2024/childrens-media-literacy-report-2024.pdf?v=368229

[18] https://www.apple.com/newsroom/2025/06/apple-expands-tools-to-help-parents-protect-kids-and-teens-online/

[19] https://blog.google/around-the-globe/google-europe/age-assurance-europe/

[20] https://www.newamerica.org/oti/briefs/exploring-privacy-preserving-age-verification/

Claims that 'age checks are best placed where content is accessed' fail to acknowledge the aforementioned issues with both privacy and efficacy. This approach has been used by the adult industry for decades now with limited effectiveness.

Should there be an overriding concern that the burden on parents is too great with respect to configuring devices, an appropriate approach would be to mandate default settings or age-verification at time of device purchase.

Success ultimately depends on major providers like Apple, Google, and Microsoft being incentivised to act. These firms control the devices, platforms, and operating systems through which enforcement can effectively occur. These firms have large and dedicated teams who genuinely care about child safety issues, but external incentives such as regulatory/political pressure and market demand for protective features can help ensure these teams have adequate internal leverage[21][22].

Content Moderation (Private Content)

Direct messages (DMs) on platforms are a source of significant harms, potentially facilitating grooming of children and cyberbullying out of public view. While platforms have some ability to detect and intervene, DMs may also be used to lure participants away from major platforms and to riskier online environments.

There is also a tension between the need to ensure that social media platforms themselves can't tamper with or surveil users' private messages, while still ensuring that there are adequate tools to protect vulnerable people.

To ensure the former, platforms have largely adopted end-to-end encryption, which, when implemented properly, ensures that a user's device is not performing the surveillance that the cryptography is designed to prevent[23].

Researchers in industry and academia have recently developed bodies of work[24] attempting to understand the extent to which it is possible to perform content moderation while preserving guarantees very similar to those of end-to-encryption. However, as of writing, none of those proposals have withstood sufficient scrutiny[25] to be considered adequate for deployment.

---

[21] There exist a variety of proposals for how to ensure effective regulatory design that attempt to maximize effectiveness by straddling direct and self-regulation, see https://www.sciencedirect.com/science/article/pii/S0267364921000303

[22] See https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf detailing the relevant internal dynamics

[23] Client-side scanning systems have been proposed and critiqued whereby instead of a third-party being able to read the messages by viewing the transmitted messages, they instead read the messages after they are arrived and decrypted. While not a formal "break" of the cryptography, it undermines the core design philosophy. For a full discussion see https://academic.oup.com/cybersecurity/article/10/1/tyad020/7590463

[24] See this review https://petsymposium.org/popets/2023/popets-2023-0060.pdf written by Prof. Scheffler

[25] https://www.wired.com/story/apple-photo-scanning-csam-communication-safety-messages/

This leaves several viable effective options as of writing.

- Platforms use *metadata* to block suspicious communications between overage and underage users, flag overage users who try to reach out to a large number of underage users, flag users who are reported by other users.
- Reporting mechanisms can reveal encrypted content to a platform operator *after* a party reports it–enabling a person to verifiably report abuse to authorities.
- Fully on-device moderation methods can "warn" the user about specific content on-device or "blur" specific contents, without sending an external signal on that detection that would break end-to-end encryption.[26]
- Platforms can segment the availability of end-to-encrypted messaging (or other features) by age[27], and then perform standard algorithmic moderation on unencrypted messages–noting that this still has significant implications for the privacy of teens.

Platform Design and Public Content Moderation

Young people are especially vulnerable to manipulative design patterns intended to drive excessive engagement or purchasing. While these patterns affect all users, youth are particularly susceptible to tactics such as intermittent reinforcement and infinite scrolling or autoplay.

Interventions for such issues should focus on regulating harmful design features. We refer the inquiry to the comments from Georgetown University's Centre for Digital Democracy's whitepaper[28], which goes into significant (but policy-helpful) detail as to particular design patterns at issue, along with potential remedies.

Existing global attempts in this area have included the UK Children's Code[29], and the ill-fated California Age-Appropriate Design code.

---

[26] For example, in Apple Messages, setting the Sensitive Content filter will use an on-device AI to detect nude images in Messages even with end-to-end encryption (without sending any notification to Apple): https://lifehacker.com/ios-17-can-blur-nsfw-pics-on-your-iphone-1850852496

[27] This does not inherently require strong age-assurance to protect young people, so long as network effects are encouraging young people to use the targeted segment (eg; Teen Accounts on Instagram https://about.instagram.com/blog/announcements/instagram-teen-accounts)–this is only possible in the absence of wholesale age-assurance restrictions which naturally push young people to attempt to access unmoderated adult versions of the platforms.

[28] https://fairplayforkids.org/wp-content/uploads/2021/05/darkpatterns.pdf

[29] https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/

Regulating design does present pitfalls, including implementation costs for small services and the risk of overbroad rules[30]. However, if effectively managed, such regulation could be effective and protective policy tool.

Content specific duties are likely a less effective policy approach. Major American platforms generally resist such duties and, when forced to comply, often restrict takedowns to local visibility rather than global removal.[31] Content based approaches are also subject to intense political contestation, domestically and internationally.

Further, for the most harmful forms of content, platforms already have strong financial interests to act proactively.

While exceptions to this analysis exist, they are limited; thus, approaches targeting specific content types should be pursued with caution.

Final Comments

The policy and technology landscape around online safety is increasingly complex. New Zealand should take time to examine each issue individually, and in sufficient technical detail, to fully understand the regulatory and technical implications of proposed interventions.

New Zealand should also be acutely aware of international political dynamics around content regulation. The U.S. has shown willingness to impose trade barriers in response to regulations it views as excessive[32]. Large platforms have responded with diplomatic pressure, withdrawal from jurisdictions, or adversarial actions. However, platforms have sometimes responded positively to well-aligned policy interventions. These dynamics, especially regarding major platforms, must be carefully considered and 'war-gamed' to realistically assess the likelihood of effective enforcement after the passage of a new rule.

Recommendations

We welcome the opportunity to provide recommendations. We also ask to be included in later stages of this or related inquiries so that we can provide more concrete and detailed recommendations as the scope of inquiry or policy reform becomes clearer.

- New Zealand should *not* adopt similar legislation to either the Online Safety Act or Australia's Social Media Minimum Age Act–the trade-offs are too significant to make. Rather, New Zealand should monitor the impacts of both on meaningful child safety outcomes, and on impacts to civil liberties, and defer consideration of such a policy under the net impact is clearer.
- Should New Zealand decide to pursue age restrictions:

[30] The California Age-Appropriate Design Code Act suffered from a number of these issues, along with constitutional infirmities, and is instructive as to patterns to avoid.
[31] https://www.reuters.com/article/business/facebook-report-shows-slight-rise-in-government-requests-for-data-idUSL3N0WI2CQ/
[32] https://www.reuters.com/world/us/trump-says-he-will-impose-retaliatory-tariffs-digital-taxes-may-come-friday-2025-02-21/

    1) the applicability should be narrowly scoped to prevent chilling speech
    2) the methods chosen for assurance should be privacy preserving
    3) the methods chosen should ensure security for users' data
    4) device-level assurance should be favoured, though accompanying rules should assure that age-assurance is not used as grounds for competition harming design (even where this arguably improves the level of assurance)
- New Zealand should not adopt policies that undermine the privacy of direct messages or other systems where users have a reasonable expectation of privacy.
- New Zealand should encourage or require the development of industry standards for trust and safety among major online platforms popular with young users (such as Instagram, TikTok, and Roblox). These standards should include transparency measures and access programs for researchers to underlying platform data.
- Where platforms offer end-to-end encrypted messaging, New Zealand:
    - Should not require platforms to provide a mechanism to reveal or scan content without a report from a user
    - Should mandate that such systems include a reporting mechanism.
- New Zealand should consider requiring platforms to segment their user base by age.
- New Zealand should fund and encourage further research, both industrial and academic, on platform safety, and on mechanisms to provide additional assurance where end-to-end encryption is used.
- New Zealand should strengthen both privacy protections for children and consider direct regulation of dark patterns

Sincerely,

Shaanan Cohney (primary author)

Sarah Scheffler

Madelyne Xiao

Lena Habtu

Appendix A: Biographical Details of Authors

Dr. Shaanan Cohney is a Senior Lecturer and the Deputy Head of School (Academic Strategy) in the School of Computing and Information Systems at the University of Melbourne. He specializes in technical and legal research at the intersection of public

policy and computer systems. He has won numerous accolades for research, including multiple Distinguished/Best paper awards in the flagship research venues in cybersecurity, and the faculty Excellence in Mid-Career Research Award. He previously served as a policy fellow in the U.S. Senate, and as the inaugural Geller Fellow placed at the U.S Federal Trade Commission. Prior to his current role he served as a Research Fellow at Princeton University's Center for Information Technology Policy.

Prof. Sarah Scheffler is an assistant professor at CyLab, the Carnegie Mellon University Security and Privacy Institute, and is faculty in the Department of Software and Societal Systems, and the Engineering and Public Policy Department. She is an applied cryptographer working at the intersection of cryptography, privacy, policy, and law. Her interdisciplinary work includes policy and technical analysis of end-to-end encryption, content moderation, compelled decryption, and privacy policies. Prior to joining CMU, Scheffler obtained her Ph.D. from Boston University in 2021 and completed postdoctoral positions at Princeton University's Center for Information Technology Policy, and MIT's Internet Policy Research Initiative.

Madelyne Xiao is a PhD candidate in the Center for Information Technology Policy at Princeton University. A former journalist, she studies content moderation, online speech, and privacy. She is currently a visiting researcher at CyLab, the Carnegie Mellon University Security and Privacy Institute. She received undergraduate degrees in mathematics and comparative literature from Stanford University before joining the editorial staff of The New Yorker in 2019.

Lena Habtu studies Political Sociology and Computer Science at Swarthmore College. She researches the governance of biometric data collection involving vulnerable populations, including age-assurance legislation as well as its collection in humanitarian contexts. Her recent research on online safety was conducted at the University of Melbourne under the Consortium of Advanced Studies Abroad (CASA).