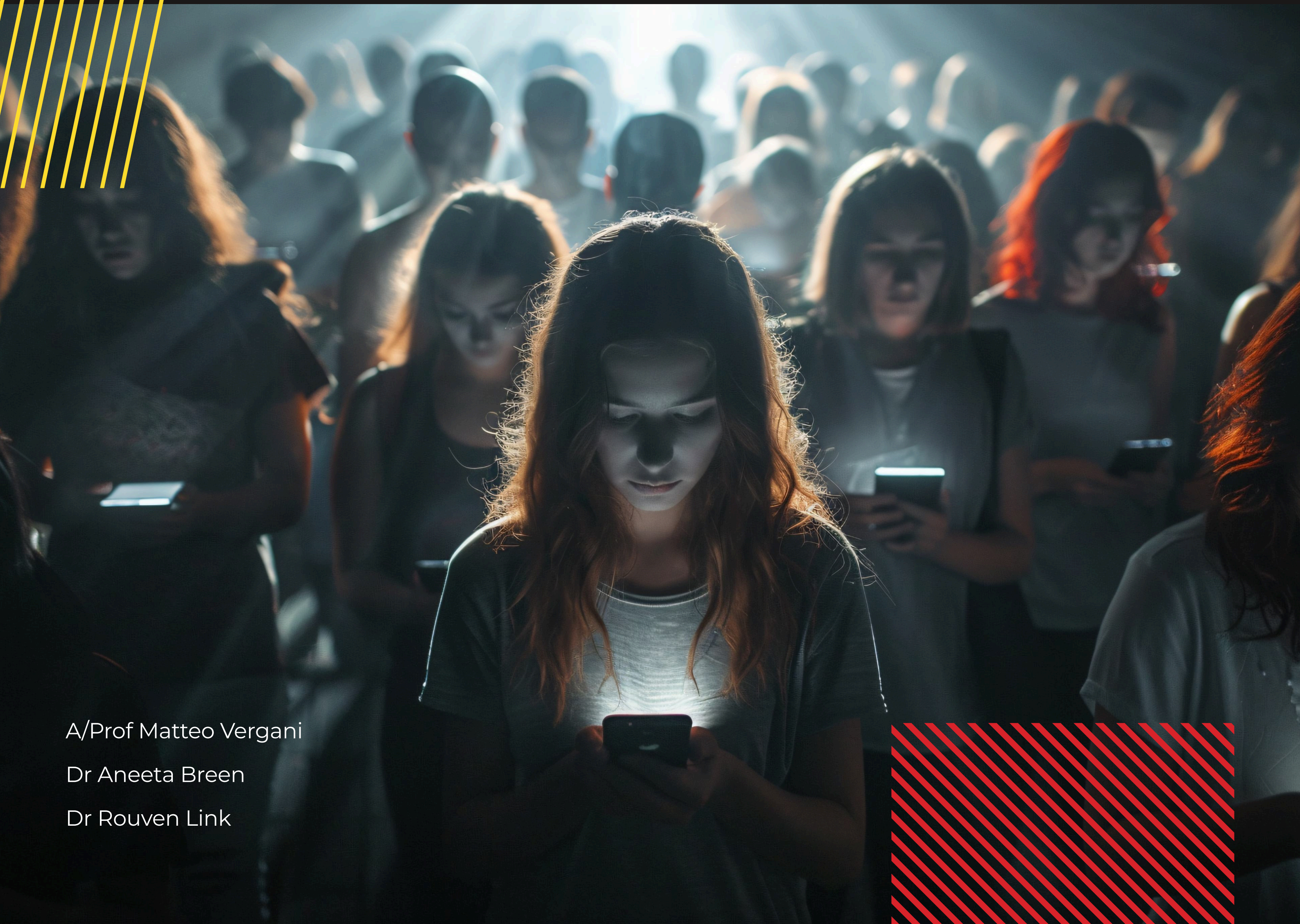
TACKLING HATE

BEYOND LAWS:

Regulating Online Hate Through Collective Action. A Map of Non-Government Strategies

Tackling Hate policy brief



A/Prof Matteo Vergani

Dr Aneeta Breen

Dr Rouven Link

INTRODUCTION

Online hate is a growing issue in Australia. According to the eSafety Commissioner (2025), one in three adults (34%) have seen online hate in the past year, and nearly one in five (18%) have personally experienced it. Among people aged 18 to 34, the rates are even higher: half (49%) have encountered online hate, and one in four (26%) have been targeted (eSafety Commissioner, 2025). The impacts can be serious: harming mental health, silencing voices, and deepening social divides.

Governments are often expected to lead the response, using civil and criminal laws to regulate online content. But legislation alone cannot solve the problem. Hate is often expressed in ambiguous or covert ways (for example, through sarcasm, humour, or coded language) that do not meet the threshold for legal action. People can intimidate others while staying just within legal limits. This makes online hate difficult to prevent and likely impossible to eliminate entirely.

That is why non-governmental attempts to regulate online hate matters. These approaches are essential and should work alongside government action. They can operate independently or in partnership with public institutions and take many different forms: from the self-regulation of social media platforms to co-regulatory approaches; from education campaigns that explain the laws, how to report hate, and what protections are available, to community initiatives and platform moderation efforts.

This policy brief attempts to map such non-governmental regulatory regulation approaches. We reviewed academic and grey literature—including peer-reviewed articles, government reports, and submissions to government inquiries—published between 2013 and 2023 in the Australian context. From this review, we identified a wide range of non-legal policy strategies and, in this report, we present a map of ideas designed as a practical resource for technology companies, non-governmental organisations (NGOs), researchers, policymakers, and others working in this space. It outlines what has been proposed, what is possible, and where future efforts could be directed.

This policy brief offers a solid starting point for anyone looking to understand or evaluate the range of ideas already on the table to counter online hate in Australia—beyond legal reforms alone.

KEY FINDINGS

We included a total of 136 documents in our review (see the methodological note for details on how they were selected). Through the analysis, we identified three main types of initiatives: **(1) self-regulation initiatives, (2) co-regulation initiatives, and (3) civil society initiatives.** Each of these approaches reflects different strategies and priorities for addressing harmful online content. In this section, we summarize the main findings and map the main Non-government approaches to regulate online hate.

01 SELF REGULATION

Self-regulation is an overarching term generally referring to 'planning and policy making regarding issues and activities not covered by public regulation' (Wotruba, 1997). This form of non-legislative regulation arises when a company, industry, or professional network establishes its own standards to govern its behaviour. Self-regulation is to be distinguished from standards and behaviours that are imposed on organisations by the state, by way of laws and regulations.

Self-regulation can be implemented, for example, through and within platforms, websites,[1] applications and other services run by industry, such as television and radio stations. Examples of online platforms that could deal with, at least, in the first instance, within their own platform, complaints about racism, include Google and Facebook. Such private entities and platforms that 'host or index online material, are referred to as 'intermediaries' by Mason and Czapski (2017, p.287). These private entities are 'intermediaries' because they do not necessarily engage in the racism themselves but racism occurs through their platforms by actual end-user offenders. Mason and Czapski's (2017) definition of intermediaries would also include industry players that are more removed from the racism, such as Internet Service Providers (ISPs) (it is also worth noting that web hosting is often handled by dedicated providers separate from ISPs, although ISPs sometimes also provide hosting services). Most of the self-regulation mechanisms take place through voluntary codes of conduct, which set the standards of behaviour that intermediaries have for users of their service. These self-regulatory mechanisms are generally derived by intermediaries or industry themselves, and include, for example, terms of service. Such voluntary in-platform self-regulation often outlines how racist content can be reported and dealt with at the industry/platform level, without resort to legal and external intervention/ escalation.

We identified 23 documents for regulating online hate through self-regulation: 11 from grey literature documents, and 12 from academic literature documents. Across these documents, four overarching themes predominate: content moderation, algorithmic regulation, accountability and transparency.

[1] 'Website' in this work is used interchangeably with 'platforms' and other services that require an Internet connection, such as applications.

Self-regulation theme 1

Content moderation

Content moderation is a self-regulatory mode of addressing harmful online material as it is essentially an internal vetting process that providers can use to publish material. For example, a media publisher, such as 'The Age' could delay publication of comments under its online articles, until it has vetted the commentary for racist or other offensive speech. Choosing to publish racist comments after vetting has been endorsed by Australian courts as making a publisher liable for racial vilification (Clarke v Nationwide News Pty Ltd, 2012). It is important to note that industry policies change over time: for example, following Elon Musk's acquisition of Twitter (now X), content moderation has been significantly downsized.

Self-regulation theme 2

Algorithmic regulation

Algorithmic regulation is another key theme emerging from our review. Several documents highlight the role of algorithmic control, for example in preventing algorithms from promoting hateful or conspiratorial content to users who have shown initial interest in such material (e.g., Matamoros-Fernández, 2017; Jakubowicz, 2018; Piracha et al., 2019). Recommendations emphasise proactive audits of algorithms that control content spread, especially to mitigate harmful impacts on minority or vulnerable communities. Such measures aim to limit 'echo chambers' that amplify harmful ideologies, thus reducing the risk of online environments becoming breeding grounds for racism and hate. However, self-regulatory mechanisms have been criticised for lack of transparency and consistency, especially regarding user anonymity (Wahlström & Törnberg, 2021).

Self-regulation theme 3

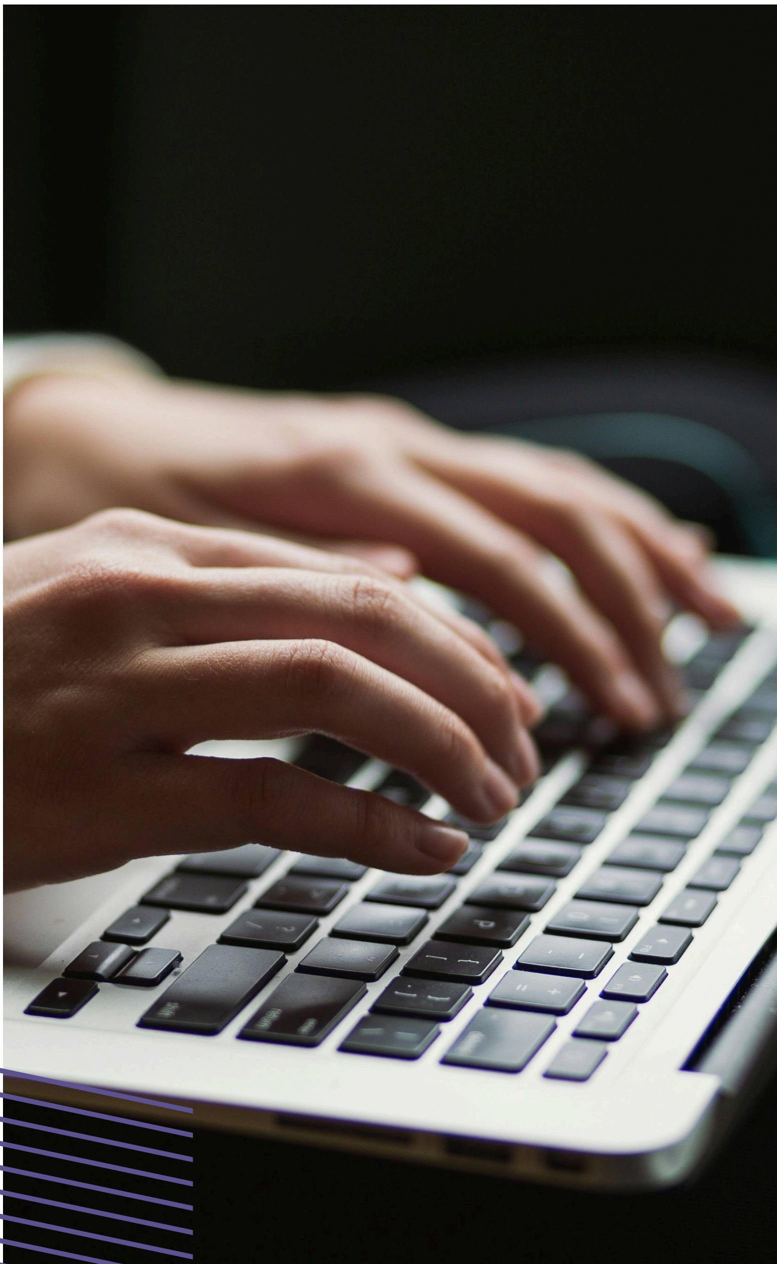
Accountability and transparency

Accountability in self-regulation is strengthened through transparency mechanisms. Numerous documents call for standardised processes to ensure platforms consistently record and address harmful incidents (e.g., Piracha et al., 2019). Public-facing reports should disclose how platforms log, archive, and resolve incidents. Regular transparency reports, disaggregated by incident type, have been proposed to give users and regulators insight into platform practices and response rates (Jakubowicz et al., 2017). Nestorovska (2022) supports quasi-regulatory mechanisms, such as Facebook's Oversight Board, as models of for independent review.

02 CO-REGULATION

Co-regulation refers to a shared approach to regulatory enforcement where both government and industry collaboratively create and uphold standards and norms (Hirsch, 2010). This could happen through each party sharing tasks, or more commonly, by operating in a graduated manner, where industry is left alone to enforce standards and laws but where it fails, state actors intervene. A notable example of co-regulation in Australia is the federal cyber-bullying regime, now governed by the Online Safety Act 2021 (Cth) (OSA), which aims to protect Australians from online harm by holding online platforms accountable for addressing harmful content swiftly (Mason & Czapski, 2017). Originally addressed under the now repealed Enhancing Online Safety Act 2015 (Cth), which focused on protecting children, the OSA has now expanded protections to all Australians, requiring online service providers to act on harmful content generally within 24 hours of receiving a complaint. Under the OSA, complaints of cyber-abuse must first be directed to the platform involved. If the platform fails to respond, the eSafety Commissioner may issue a formal removal notice, creating a structured escalation model.

We identified 20 documents proposing to regulate online hate through co-regulation: 12 from grey literature documents, and eight from academic literature documents. Across these documents, four overarching themes predominate: algorithmic transparency and accountability oversight, taking down hateful content within specific timeframes, mandatory users reporting, and a tiered redress mechanism.

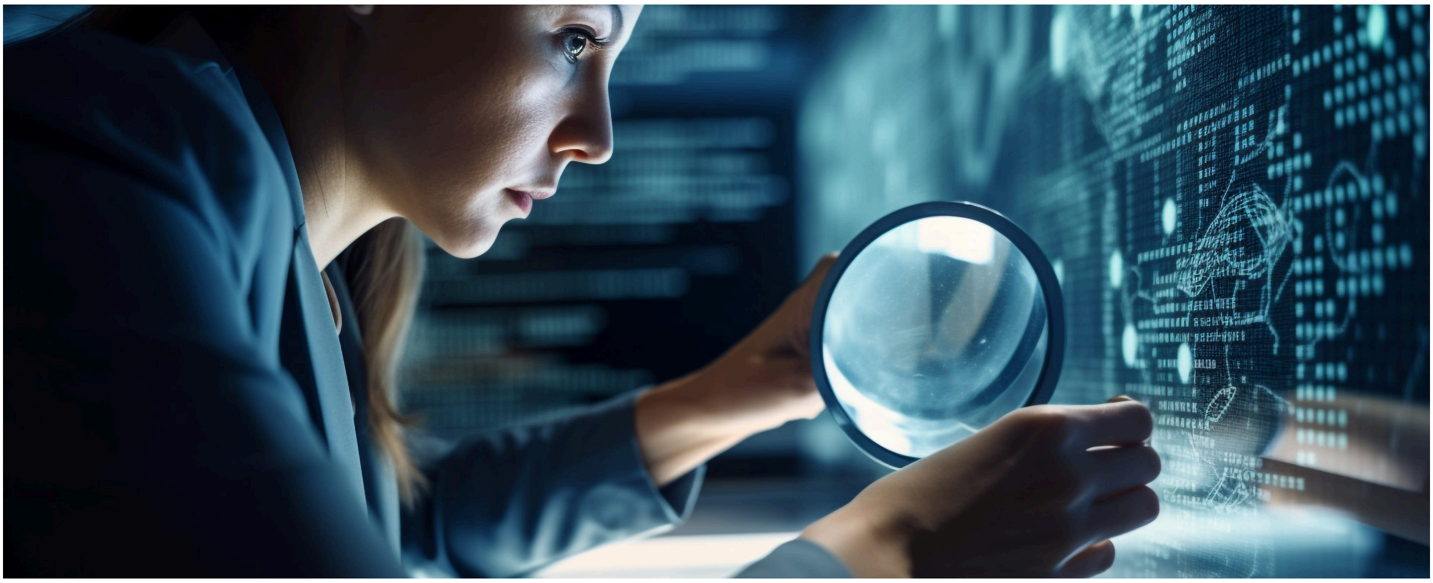


Co-regulation theme 1

Mandatory reporting standards

The first co-regulation theme is the mandate for regular audits, detailed transparency reports, and penalties for inaction. Many documents suggest that independent regulators should conduct audits, focusing on algorithmic and moderation outcomes rather than the mechanisms themselves. For example, platforms might be required to disclose the logic behind algorithmic decisions or allow external regulators access to data to assess relevant societal impacts (Taylor, 2020).

By focusing on outcomes rather than proprietary code, regulators can hold platforms accountable without compromising intellectual property or corporate secrets (Oboler & Connelly, 2018). In relation to moderation, a quality control process to review moderation decisions with a unique ID system for tracking complaints would help build public trust in online platforms' commitment to managing hate speech.



Algorithmic transparency is largely subject to self-regulation in Australia, through the Australian Code of Practice on Disinformation and Misinformation (the Code), which was developed by the Digital Industry Group Inc. and published in 2021, with guidance provided by the Australian Communications and Media Authority (ACMA). The Code requires digital platforms to minimise harm emerging from misinformation and disinformation on their own platforms. While the Code is technically a self-regulatory measure, the ACMA does monitor its operation and effectiveness, which gives it a co-regulatory flavour, in essence. The ACMA, as the Australian communications and media services regulator, has reported that signatories to the Code have been ineffective in detailing exactly how they will combat disinformation. To improve such and other efforts, the federal Australian government is currently publicly consulting on a bill that could give the ACMA formal powers of co-regulation, where platforms fail to meet their obligations under the Code.

Co-regulation theme 2

Taking down hateful content within specific timeframes

Platforms could be mandated to take down hateful content within specified timeframes, as in the German model, where platforms must remove illegal content within 24 hours to avoid penalties (Jakubowicz et al., 2017). This approach is designed to motivate platforms to take consistent and meaningful action on harmful content, ensuring they cannot evade responsibility without legal consequences, especially when repeated breaches occur.

Co-regulation theme 3

Mandatory reporting of hateful users

Many documents advocate for a requirement for social media platforms themselves, to report users who repeatedly engage in hate speech to authorities. This reporting would help law enforcement identify and track persistent offenders, potentially leading to prosecution for ongoing online hate. Platforms could be tasked with recording data such as IP addresses and content logs, which could be forwarded to authorities once users repeatedly breach standards. This escalation model would activate automatically when certain thresholds are met, ensuring a structured response that deters repeat offenders. This approach requires effective collaboration mechanisms between platforms, regulatory agencies, and law enforcement.

Co-regulation theme 4

Tiered redress mechanisms

Many documents endorse a tiered redress mechanism that starts with user-initiated reports on platforms, with a pathway to escalate to external regulatory intervention if internal responses are insufficient. This model aims to make it easier for users to report offensive content directly to the platform while providing an external appeals process if platforms fail to act. The scheme's strength lies in empowering users to report hate speech and discrimination, reducing barriers to initial complaint filing. If platform actions are inadequate, external bodies like the eSafety Commissioner could intervene, enforcing penalties for non-compliance or escalating unresolved cases to legal authorities (Flew & Gillet, 2021). Such models draw on the success of Australia's cyber-bullying laws, which combine platform responsibility with governmental oversight, illustrating a feasible pathway to applying similar principles to online hate regulation (Nestorovska, 2022).

A tiered or graduated redress mechanism for some content that could be classified as online hate speech does exist in a limited capacity under the cyber-abuse regime administered by the eSafety Commissioner. The current regime requires complainants to contact platforms first, and if the platforms do not assist, then the eSafety Commissioner can mandate the removal of seriously harmful material. As noted earlier, however, the cyber-abuse regime administered by the eSafety Commissioner operates only for directly affected (individual) adults or children and not for collective hate on the basis of a group characteristic.

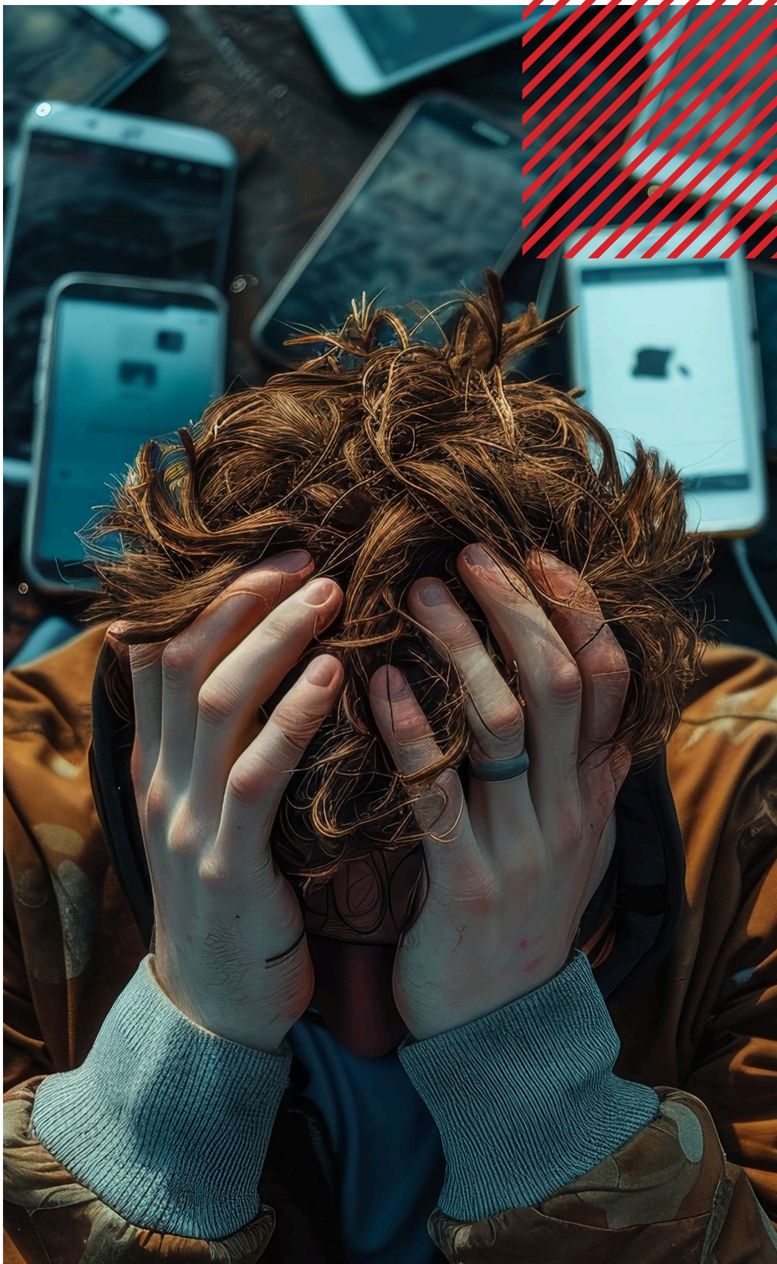
The extent of government intervention

An important difference between the above co-regulation approaches lies in the proposed extent of government intervention. As noted above, some documents call for strong regulatory oversight, including mandated audits and penalties, with the eSafety Commissioner taking a central role in ensuring compliance. Others suggest a graduated response model, where platforms initially manage content moderation but escalate unresolved cases to authorities, adapting the cyber-bullying framework to address online hate. Similarly, some documents advocate for mandatory algorithmic audits and transparency regarding content amplification, while others propose that cooperation between stakeholders, including civil society organizations and government agencies, is essential to addressing online hate effectively without solely relying on direct access to platform algorithms. Some call for extensive, regular disclosures by platforms on moderation practices, enabling civil society and policymakers to assess platform compliance, while others argue for more selective reporting to avoid imposing excessive burdens on platforms.

03 CIVIL SOCIETY INITIATIVES

Civil society regulation includes strategies where groups such as community organisations, NGOs, and associations operate in different ways to support, or take part in legal or regulatory measures. This approach leverages civil society's unique proximity to affected communities, enabling responsive and adaptable interventions that complement broader legislative and policy frameworks.

We identified 50 documents for regulating online hate through civil society regulation: 31 from grey literature documents, and 19 from academic literature documents. Across these documents, four key themes emerged: collaboration between platforms and civil society, community-led reporting and data collection mechanisms, shifting societal norms, support networks and resources for affected communities.



Civil society regulation theme 1

Collaboration between platforms and civil society

Many documents proposed that platforms should work with civil society groups, for example anti-racism organisations, to address community-specific hate speech and improve responsiveness to localised issues (Jakubowicz et al., 2017). This model suggests that civil society organisations could act as third-party flaggers or advisers on hate speech standards, helping platforms align their moderation efforts with community values. The establishment of a permanent taskforce or partnership network between social media platforms, civil society organisations and local government bodies could be a solution, ensuring continuous dialogue and cooperative problem-solving. This inclusive approach aims to build resilience against online hate by combining the technical expertise of platforms, the enforcement power of government, and the grassroots insight of civil society (Oboler et al., 2019).

Civil society regulation theme 2

Community-led reporting and data collection mechanisms

A key proposal emerging from the analyses is to support community-led reporting and data collection mechanisms, which are widely advocated as tools to empower communities in documenting and responding to online hate incidents. Such mechanisms provide safe and trusted channels for individuals to report incidents of hate without relying solely on formal law enforcement, which some communities may mistrust due to past negative experiences. Civil society led programs like Call It Out, the Islamophobia Register Australia, or the Antisemitism Report, demonstrate how community-based reporting platforms can foster localised data collection while facilitating connections to authorities for further action. Initiatives like the Online Hate Prevention Institute’s "Fight Against Hate" software further enhance data collection by allowing users to report and categorise hate incidents on social media, which contributes to a larger dataset on digital racism in Australia (Oboler & Connelly, 2018). Such community-led reporting tools address gaps in formal reporting, amplify marginalised voices, and provide empirical evidence to inform anti-racism policies and public awareness campaigns.

Civil society regulation theme 3

Shifting societal norms

Many of the documents we reviewed discussed efforts to shift societal norms about what constitutes admissible and inadmissible speech in online environments. These efforts are often supported by governments and technology companies through the sponsorship of civil society advocacy. Norms can be shifted in several ways. The first is through education campaigns and training, such as anti-racism programs. For example, Be Deadly Online and Codey, developed for First Nations youth, provide culturally tailored resources to help young people identify and respond to racism they encounter online. These initiatives demonstrate how digital safety education can be combined with strategies to address cyber-racism (Montgomery, 2014).

A second approach is advocacy and leadership from political and public figures. Public statements from politicians and community leaders that explicitly reject hate speech can help foster a culture that does not tolerate online hate and racism. A third approach involves counter-speech and positive narrative campaigns, which challenge hate speech directly while promoting inclusive values.

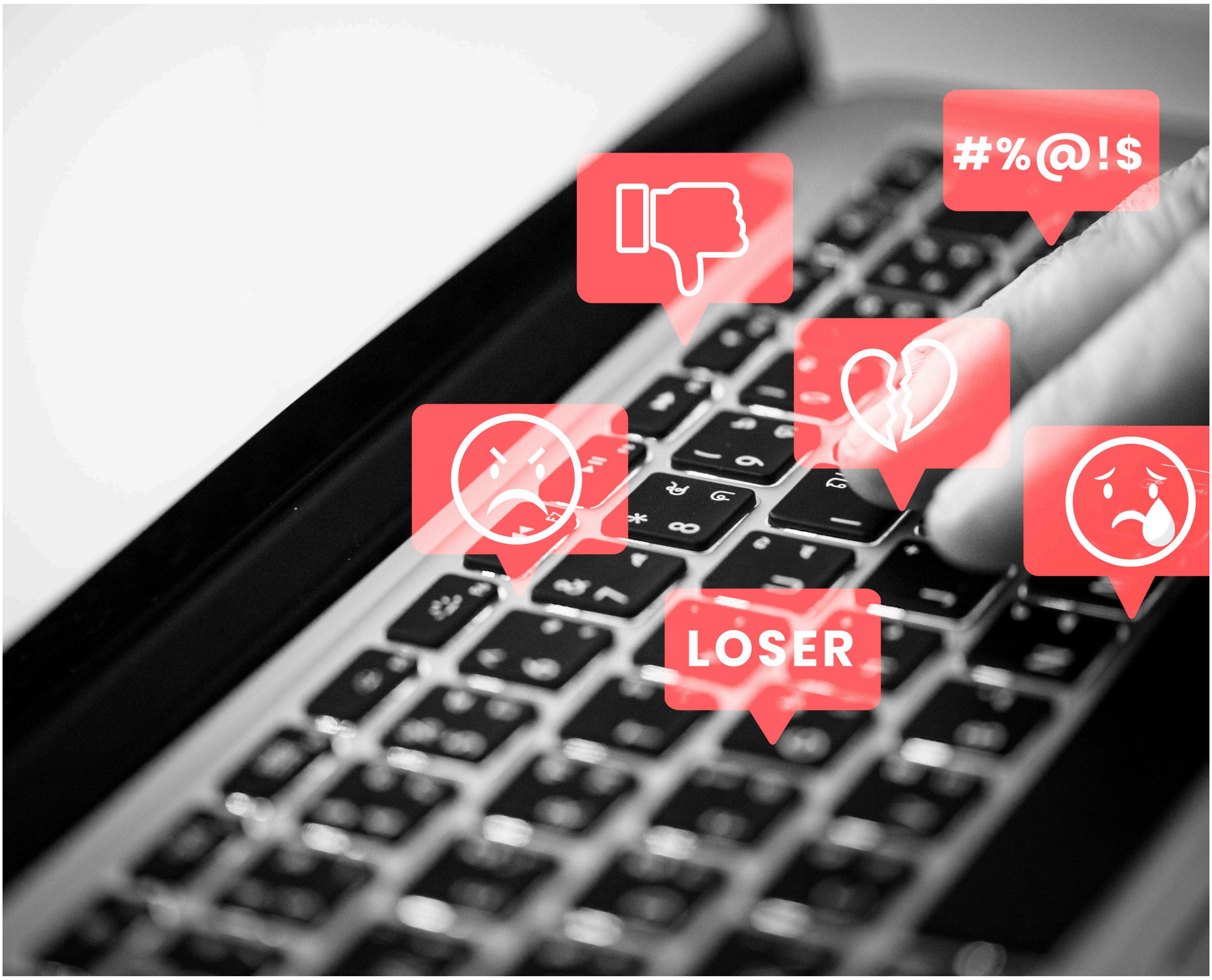
Civil society regulation theme 4

Support networks and resources for affected communities

Many documents recommended resourcing support networks and community structures, recognising them as essential for those affected by online hate. Such initiatives help address immediate needs while also fostering long-term resilience. Localised support structures—such as peer networks and mental health services—are increasingly promoted for communities experiencing high levels of digital harassment, as they provide both psychological and social support (Johnson & West, 2023). One example is the Islamophobia Support and Outreach project, which highlights the value of community-controlled platforms. By catering to specific cultural or religious needs, such initiatives can offer safety, empowerment, and a stronger sense of belonging for community members.

The limitations of civil society regulation

Civil society regulation faces several limitations. Educational initiatives and counter-speech campaigns, though valuable, are often criticised for their limited impact when unsupported by enforceable standards or systemic reforms to address underlying discrimination (Jakubowicz, 2017; Gelber, 2021). Moreover, there is little consensus on the appropriate role of civil society: some advocate for organisations to act as third-party flaggers of harmful content (Jakubowicz et al., 2017), while others see their role as primarily supportive, focusing on public education rather than direct moderation (Flew & Gillet, 2021). While community-led reporting systems are sometimes promoted as effective tools, others suggest that their close association with government agencies (in particular law enforcement) can deter participation from groups with historical mistrust of authorities. These tensions highlight the difficulties of designing a co-regulatory framework that leverages civil society's strengths while ensuring accountability and effectiveness in addressing online hate.



CONCLUSION

The mapping of non-government strategies to regulating online hate offers an understanding of the range and scope of tools to address online hate, with the aim to facilitate collaboration among governments, digital platforms, community organisations, and other stakeholders.

For policymakers, this mapping work serves as a reference tool to identify effective strategies and approaches for collaborating with non-government actors that can be adapted or adopted in their respective jurisdictions.

For stakeholders, including advocacy groups and community organisations, the mapping exercise provides a foundation for targeted advocacy and program design. By presenting this information in a structured way, the mapping exercise also facilitates collaboration and dialogue among stakeholders, ensuring that efforts to address online racism and hate are coordinated and impactful.

Scientific evidence suggests that non-government approaches to regulating hate might have the potential to work. For example, Álvarez-Benjumea and Winter (2018) found that removing hate speech can reduce its occurrence by reinforcing descriptive norms over injunctive norms. Similarly, Yildirim et al. (2023) demonstrated that issuing warnings about potential suspensions on Twitter resulted in a 10% reduction in hate speech, with the most effective warnings being those perceived as legitimate. Although these findings are based on specific contexts, and the effects of broader application remain inconclusive, they suggest that content moderation enforced by platforms can have a strong impact on reducing hate speech and changing social norms.

Co-regulatory approaches have been implemented in various jurisdictions. The European Union Code of Conduct on countering illegal hate speech demonstrates how shared accountability can improve responsiveness to harmful content, requiring social media companies to act on hate speech reports within 24 hours. However, its effectiveness in the Australian context requires further examination, as the success of co-regulation depends on the adaptability of frameworks to local legal, cultural, and technological environments.

Civil society initiatives have also contributed significantly, particularly through education and counter-speech campaigns, to reinforce regulatory frameworks. For example, Wachs et al. (2020) evaluated Germany's "HateLess" initiative, finding that it reduced hate speech perpetration and victimisation among adolescents, partly by increasing empathy and self-efficacy. Similarly, the "Kombat with Kindness" campaign in Utah demonstrated reductions in students' exposure to hate messages through peer-led activities (Savoia et al., 2019). Other initiatives, such as targeted social media campaigns in Indonesia, have shown modest reductions in online hate speech, although their long-term effectiveness remains unclear (Bodine-Baron et al., 2020).

Overall, while various approaches show potential, the evidence remains fragmented and insufficient for drawing definitive conclusions. A more robust and systematic evaluation is necessary to determine the most effective strategies for addressing online racism and hate in Australia and beyond. Future efforts should prioritise comprehensive, longitudinal studies that consider the interplay of regulatory, technological, and social factors in shaping online behaviours.

Methodological Note

This policy brief is based on a systematic review of academic and grey literature published in English between 2013 and 2023. We included literature that discussed the regulation of online racism and hate in the Australian context.

Academic literature was identified using a structured search strategy across 22 databases and citation indices. Four key studies were used to benchmark the search and screening process, and a final search string combining 14 terms returned 2,032 records. After removing duplicates and applying inclusion criteria (e.g. publication year, Australian focus, relevance to hate or racism in the media), 1,642 records underwent a two-stage screening using EPPI Reviewer 6. The first stage assessed titles and abstracts for relevance; the second involved full-text screening to identify specific, detailed regulatory documents. A total of 37 academic documents were included in the final analysis.

Grey literature was sourced through manual searches of government and NGO websites, yielding 15,569 documents. After de-duplication, OpenAI's ChatGPT 4.0o was used to assist in two-stage pre-screening based on thematic relevance and the presence of detailed proposals. This narrowed the selection to 179 documents, which were then manually reviewed in EPPI Reviewer 6, resulting in 99 grey literature documents being included.

Data extraction and analysis were conducted by a single reviewer. Each proposal identified was categorised into one of five regulatory types: self-regulation, legal regulation (criminal), legal regulation (civil), co-regulation, and civil society regulation. A thematic analysis of these documents informed the findings presented in this policy brief.

References

Álvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. European Sociological Review, 34(3), 223–237.

Aroney, N., & Taylor, P. (2023). Building tolerance into hate speech laws: State and territory anti-vilification legislation reviewed against international law standards. University of Queensland Law Journal, 42(3), 317–346. https://doi.org/10.38127/uqlj.v42i3.8537

Bodine-Baron, E., Marrone, J. V., Helmus, T. C., & Schlang, D. (2020). Countering Violent Extremism in Indonesia.

Clarke v Nationwide News Pty Ltd (2012) 201 FCR 389 (Austl.).

eSafety Commissioner. (2025). Hate in the digital age: Adults' encounters with online hate. Australian Government. https://www.esafety.gov.au/research/encounters-with-online-hate#download-the-reports-and-methodology

Flew, T., & Gillet, R. (2021). Platform policy: Evaluating different responses to the challenges of platform power. Journal of Digital Media & Policy, 12(2), 231–246. https://doi.org/10.1386/jdmp_00061_1

Gelber, K. (2021a). Differentiating hate speech: A systemic discrimination approach. Critical Review of International Social & Political Philosophy, 24(4), 393–414. https://doi.org/10.1080/13698230.2019.1576006

Hirsch, D. D. (2010). The law and policy of online privacy: Regulation, self-regulation, or co-regulation. Seattle University Law Review, 34, 439.'

Jakubowicz, A. (2017). Alt_Right White Lite: Trolling, hate speech and cyber racism on social media. Cosmopolitan Civil Societies: An Interdisciplinary Journal, 9(3), 41–60. https://doi.org/10.5130/ccs.v9i3.5655

Jakubowicz, A. (2018). Algorithms of hate: How the internet facilitates the spread of racism and how public policy might help stem the impact'. Journal and Proceedings of the Royal Society of New South Wales, 151(1): 69–81.

Johnson, B., & West, R. (2023). Ableism versus free speech in Australia: challenging online hate speech toward people with Down syndrome. Disability & society, 38(9), 1711-1733.

Mason, G., & Czapski, N. (2017). Regulating cyber-racism. Melbourne University Law Review, 41(1), 284-340.

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. Information, Communication & Society, 20(6), 930–946. https://doi.org/10.1080/1369118X.2017.1293130

Nestorovska, D. (2022). Press councils: Adapting an existing self-regulatory model for the social media age. Computers and Law: Journal for the Australian and New Zealand Societies for Computers and the Law, 94.

Oboler, A., & Connelly, K. (2018). Building SMARTER Communities of resistance and solidarity. Cosmopolitan Civil Societies: An Interdisciplinary Journal, 10(2), 91–110. https://doi.org/10.5130/ccs.v10i2.6035

Piracha, A., Sharples, R., Forrest, J., & Dunn, K. (2019). Racism in the sharing economy: Regulatory challenges in a neo-liberal cyber world. Geoforum, 98, 144-152.

Savoia, E., Su, M., Harriman, N., & Testa, M. A. (2019). Evaluation of a school campaign to reduce hatred. Journal for Deradicalization, (21), 43-83.

Taylor, P. (2020). Uneven platforms: The press, social media, search engines and freedom of expression. University of Tasmania Law Review, 39(2), 121–149.

Turner, G. (2018). The media and democracy in the digital era: Is this what we had in mind? Media International Australia, 168(1), 3–14. https://doi.org/10.1177/1329878X18782987

Wachs, S., Wright, M. F., & Gámez-Guadix, M. (2024). From hate speech to HateLess. The effectiveness of a prevention program on adolescents' online hate speech involvement. Computers in Human Behavior, 157, 108250.

Wahlström, M., Törnberg, A., & Ekbrand, H. (2021). Dynamics of violent and dehumanizing rhetoric in far-right social media. New media & society, 23(11), 3290-3311.

Wotruba, T. R. (1997). Industry self-regulation: A review and extension to a global setting. Journal of Public Policy & Marketing, 16(1), 38-54.

Yildirim, M. M., Nagler, J., Bonneau, R., & Tucker, J. A. (2023). Short of suspension: How suspension warnings can reduce hate speech on Twitter. Perspectives on Politics, 21(2), 651–663.

TACKLING ~~HATE~~

A/Prof Matteo Vergani

Dr Aneeta Breen

Dr Rouven Link