WHEN SPEECH CROSSES THE LINE

The Case for a Clear Legal Threshold Between Free Expression and Online Harm

A Digital Defence Policy Paper | October 2026

Zara Cooper

Chief Executive Officer, Digital Defence

KEY ARGUMENT

There is no such thing as absolute free speech. Every democratic legal system already places limits on expression: defamation law, copyright, contempt of court, vilification law, carriage service offences, national security legislation, and electoral laws all draw lines around what can and cannot be said. The problem is not that those lines exist. The problem is that no jurisdiction has drawn them clearly enough to distinguish robust public debate from conduct that causes real harm and breaches existing law. That failure does not protect free speech. It protects those who exploit the absence of clarity to harm others while hiding behind it.

1. The myth of absolute free speech

The phrase "free speech" is invoked as though it describes an absolute right that exists somewhere above the law, untouched by regulation. It does not and never has. In every jurisdiction that claims to protect freedom of expression, including Australia, the United States, the United Kingdom, the European Union, and Canada, the law already draws limits on what can be said, published, broadcast, or posted online. Those limits are extensive, well-established, and largely uncontroversial when people encounter them in their original contexts. It is only when they are applied to online harm that the cry of "free speech" is raised, as though the internet were somehow different in kind from every other medium of human communication that law has ever had to grapple with.

Consider what expression is already lawfully restricted in Australia alone. Defamation law prevents the publication of false statements of fact that damage a person's reputation, and it applies to online posts with exactly the same force as it applies to a newspaper. Copyright law restricts the reproduction and

distribution of creative work without the rights-holder's permission, and it does so on social media platforms as readily as in any bookshop. Contempt of court prevents publication of material that could prejudice a fair trial, which is why journalists routinely hold back information that is in the public interest in order to comply with suppression orders. Carriage service offences under the *Criminal Code Act 1995* (Cth) make it an offence to use a carriage service, including a social media platform, to menace, harass, or cause offence. Racial vilification laws in Victoria, Queensland, and New South Wales make it unlawful to incite hatred against a person or group on the basis of race. National security legislation restricts disclosure of classified information, glorification of terrorism, and incitement to political violence. Electoral law restricts what can be published about candidates in the lead-up to an election and how. Privacy law restricts the collection, use, and disclosure of personal information. Sedition law, though rarely used, restricts urging force or violence against the sovereign or government.

These restrictions exist across the full breadth of democratic life. They were enacted by parliaments, upheld by courts, and are operated by regulators every day. None of them has destroyed free speech. All of them have been accepted, after sometimes vigorous public debate, as legitimate limits on expression in the public interest. The question has never been whether expression can be regulated. The question has always been where to draw the line and how.

2. The laws that already govern speech: a comprehensive picture

It is worth setting out in detail examples of existing legal constraints on expression in Australia, because the breadth of that picture tends to be underappreciated in public debates about online regulation. The following categories of law already restrict what Australians may lawfully say, write, broadcast, or post online.

Defamation: the *Defamation Act 2005* (Cth), which applies in substantially uniform form across Australian states and territories, makes it unlawful to publish material that would cause an ordinary reasonable person to think less of another person, and provides for substantial damages. Truth is a complete defence, but the threshold for publication is low: a single social media post sent to one person can constitute publication.

Copyright: the *Copyright Act 1968* (Cth) restricts the reproduction, communication, and adaptation of protected works without licence or a specific fair dealing exemption. This applies to text, images, music, and audiovisual material shared online.

Contempt of court: publishing material that scandalises the court, prejudices proceedings, or disobeys a court order or suppression notice is contempt at common law and under specific legislation, and can attract custodial sentences.

Carriage service offences: sections 474.15 to 474.17 of the *Criminal Code Act 1995* (Cth) make it a criminal offence to use a carriage service, which includes any internet-connected platform, to make a threat, to harass, menace, or cause offence, or to transmit material that a reasonable person would find menacing, harassing, or offensive. These provisions have existed for more than two decades.

Vilification laws: the *Racial Discrimination Act 1975 (Cth)* makes it unlawful to do an act, in public, that is reasonably likely to offend, insult, humiliate, or intimidate a person or group on the basis of race, colour, or national or ethnic origin. State and territory legislation extends similar protections to religion in some jurisdictions and to other attributes including gender identity and sexual orientation.

National security and counter-terrorism: the *Criminal Code Act 1995* (Cth) contains extensive provisions restricting advocacy of terrorism, support for proscribed organisations, sedition, and

disclosure of intelligence information. Section 80.2A makes it an offence to urge force or violence against a group on the basis of race, religion, nationality, national origin, or political opinion.

Electoral law: the *Electoral Act 1918* (Cth) restricts how candidates and parties may be identified in political advertising, requires authorisation of political material, and prohibits certain conduct in the lead-up to elections and on polling day.

Privacy law: the *Privacy Act 1988* (Cth) restricts the collection, use, and disclosure of personal information by entities to which it applies, including large platforms, and the Privacy and Other Legislation Amendment Act 2024 (Cth) has recently introduced doxxing offences targeting the malicious disclosure of personal information.

Consumer protection: the Australian Consumer Law prohibits misleading and deceptive conduct, which extends to false or misleading representations made in commercial contexts online.

Advertising standards: the Ad Standards framework, operated by Ad Standards Australia, restricts advertising content on grounds of taste, decency, discrimination, and harm, and applies to paid promotional content on social media platforms.

These are not fringe or controversial laws. They are the ordinary legal framework within which Australians communicate every day, largely without noticing. Every person who has been served with a defamation notice, every journalist who has held a story because of a suppression order, every company that has had an advertising complaint upheld, and every person charged under the carriage service offences has experienced the reality that free speech in Australia is already extensively regulated. **The only place that seems genuinely confused about this is the online space.**

3. The failure no jurisdiction has fixed: drawing the line

If the legal framework already regulates speech so extensively, why does the online harm problem persist? The answer is that existing law was not designed to address the volume, velocity, anonymity, and cross-jurisdictional nature of online harm, and no jurisdiction has yet drawn a clear, comprehensive line between conduct that remains within the bounds of lawful expression and conduct that crosses into actionable harm.

That absence of clarity causes two distinct problems, and they pull in opposite directions. On one side, victims of genuine harm, including people subjected to coordinated pile-on attacks, image-based abuse, sustained doxxing, and targeted vilification, cannot obtain timely redress because the legal frameworks available to them were not designed for their situation: the thresholds are unclear, the enforcement pathways are fragmented, and the burden of proof and the cost of proceedings place access to justice beyond the reach of most people. On the other side, the absence of a clear line makes the defence of "free speech" easy to weaponise: a person engaged in sustained targeted abuse of a real, identifiable individual can characterise every individual communication as an expression of opinion, and without a framework that addresses the aggregate conduct, they are often right as a matter of law.

The result is a system that fails simultaneously at both of the things it is supposed to do. It does not protect speech, because the chilling effect of unaddressed online abuse causes real people, particularly women, minorities, and people in public life, to withdraw from public discourse. It does not protect victims, because the enforcement tools available to them are inadequate to the scale and character of the harm they experience. A clear legal threshold would address both failures at once.

Australia is not alone in this. Across comparable jurisdictions, the pattern is the same. The Online Safety Act in the United Kingdom provides for content removal but does not establish a graduated enforcement framework with financial penalties. The Digital Services Act in the European Union addresses systemic platform risk but does not create a complaint-based individual enforcement mechanism with clear harm categories. The Communications Decency Act in the United States has historically shielded platforms from liability without providing any pathway for individual victims. Canada's Online Harms Act, passed in 2024, has taken steps toward harm classification but does not yet provide the civil penalty and demerit point framework that proportionate enforcement requires. None of these frameworks has drawn the clear line between discourse and harm that victims need and that genuine defenders of free expression should welcome.

4. Why clarity protects both speech and victims

There is a persistent assumption in public debate about online regulation that protecting victims necessarily comes at the cost of free speech, and that anyone who argues for greater legal clarity around online harm is therefore an opponent of free expression. That assumption is wrong, and it is worth explaining precisely why.

Clarity in law protects speech by defining what is protected. The most effective protection for free expression is not an absence of rules, but a clear statement of what is permitted. When a journalist knows exactly what the defamation threshold is, they can publish up to and including that line with confidence. When a politician knows exactly what electoral communication law requires, they can campaign freely within those requirements. When a satirist knows exactly what the legal definition of satire requires, they can write without fear. The absence of clear rules does not expand freedom: it expands uncertainty, and uncertainty chills expression at least as effectively as explicit prohibition.

The same logic applies to online harm. When a person posting about a matter of public controversy knows clearly that the legal threshold for vilification requires incitement to hatred directed at a person or group, and not merely the expression of controversial views, they can express those views without fear. When a person engaging in robust criticism of a public figure knows clearly that the threshold requires targeting, threats, or calls to harassment, they can criticise without fear. Clarity makes the space inside the line genuinely safe for expression. Ambiguity does not.

Clarity in law also protects victims by giving enforcement agencies something to enforce. The failure of existing enforcement pathways is not primarily a failure of will, it is a failure of legal architecture. Police, regulators, and courts cannot act consistently or effectively on conduct that is not clearly defined. When the law does not specify what a coordinated pile-on attack is, how it is assessed, and what threshold must be crossed before it is actionable, enforcement is inevitably inconsistent, delayed, and inadequate. Clear harm categories, clear thresholds, and clear enforcement pathways are the preconditions for timely, proportionate, and credible enforcement action.

5. How the defence of free speech is weaponised by those who breach existing law

"Free speech" as a shield for harmful online conduct is not an argument made in good faith by people genuinely concerned about government overreach. It is a tactic used by people who are engaged in conduct that already breaches existing law, and who rely on the absence of a clear, accessible, and timely enforcement mechanism to avoid accountability. Understanding how the tactic works is important to understanding why legal clarity is the response it cannot survive.

The mechanism operates as follows. A person engaged in sustained targeted abuse of an individual, or a coordinated campaign designed to drive a person from public discourse, characterises each individual communication as an expression of opinion. Any single message, viewed in isolation, may or may not meet the existing threshold for a carriage service offence. The aggregate effect of hundreds or thousands of such messages, sent by dozens of participants over days or weeks, is clearly capable of constituting serious harm and breaching existing law. But the law as currently structured addresses individual communications, not aggregate campaigns, and enforcement agencies lack both the legal architecture and the resources to address the conduct as a whole. The "free speech" argument fills the gap: it provides a rhetorical defence for conduct that, if assessed as a whole, would plainly be unlawful.

The same tactic is used in relation to targeted doxxing. Publishing a person's home address alongside calls to "make your feelings known" does not, in most Australian jurisdictions, constitute a clear criminal offence unless physical harm results. The person who does so can claim they are merely publishing information that is, in some technical sense, available in the public domain, and that calling on others to express a view is simply the exercise of free expression. The harm to the target is real, immediate, and often severe. The legal framework available to address it is inadequate. The "free speech" argument exploits that inadequacy.

A clear line between discourse and harm does not just protect victims: it exposes the bad faith of those who invoke free speech as a shield. When the law clearly defines what coordinated harassment is, when it clearly states that the aggregate effect of a campaign is the relevant measure rather than the isolated effect of each individual message, and when it provides a timely and proportionate enforcement pathway, the tactic of hiding behind individual communications becomes unavailable. The argument that each tweet was just an opinion loses its force when the framework addresses the campaign as a whole.

6. What a clear threshold looks like: the Digital Defence proposal

Digital Defence has developed a draft Harm Classification Framework as part of its Buckle Up: Road Rules for the Online Space legislative proposal. **The Framework represents a first attempt to draw the clear line that no jurisdiction has yet drawn, and to build around that line a proportionate enforcement architecture.** It rests on three principles that follow directly from the analysis above.

- **First, the threshold question must be stated explicitly in the law.** The question in every case is whether conduct has crossed from the expression of views, however strongly held, controversial, or offensive, into conduct that a reasonable person would regard as threatening, targeting, harassing, or inciting in a manner that would breach existing law or the civil threshold. That question must be the test, and it must be stated clearly and applied consistently, rather than left to be inferred case by case from fragmentary existing provisions.
- **Second, carve-outs for protected expression must be explicit and illustrative, not exhaustive.** Each harm category in the Framework includes clear examples of conduct that does not cross the threshold: robust criticism, political debate, theological disagreement, commentary on matters of public interest, satire, and academic inquiry. Those carve-outs are not an exhaustive list of protected expression: they are illustrations of how the threshold operates in practice, and they make clear that the threshold is crossed by threatening, targeting, or inciting, not by expressing.

- **Third, enforcement must be proportionate, timely, and credible.** The Framework proposes a civil penalty and demerit point system modelled on road traffic enforcement, where the financial penalty is scaled to the severity of the conduct and the demerit points accumulate with repeat conduct, escalating progressively to suspension and deplatforming. The analogy to road rules is deliberate: road rules do not prevent people from driving. They create a common framework within which driving is safe for everyone. Online conduct rules should do the same.

The Framework also draws a clear distinction between the civil and criminal thresholds. Conduct at the lower end of the harm spectrum, where the appropriate response is a fine and a warning, is addressed by the civil penalty pathway. Conduct at the upper end, including threats of violence, CSAM, and incitement to terrorist violence, routes directly to criminal enforcement. The civil pathway does not displace the criminal law: it fills the space between criminal conduct and no enforcement at all, which is currently where most online harm lives.

7. Conclusion: the line protects everyone

The debate about free speech and online harm is conducted, too often, as though it were a zero-sum contest between those who want to protect expression and those who want to restrict it. It is not. The failure to draw a clear line between discourse and harm damages both sides of that false contest. It leaves genuine victims without redress, and it leaves genuine defenders of free expression without a clear statement of what is protected.

Every democratic legal system already limits expression in dozens of ways. The question is not whether limits exist, but whether the limits that apply to online conduct are clear, proportionate, accessible, and enforced. Australia, like every comparable jurisdiction, has so far failed to answer that question adequately. The consequence is a regulatory environment in which those who cause the greatest harm are those most able to exploit the ambiguity, and those who most need the protection of a clear line are those least able to access it.

Drawing the line clearly is not a threat to free expression. It is the precondition for it. People speak freely when they know what is protected. People are protected from harm when the law is clear about what crosses the line. The two goals are not in tension. They are the same goal, pursued together through the same instrument: a clear, comprehensive, and credible legal threshold between speech and the conduct that exploits its absence.

About Digital Defence

Digital Defence is an independent organisation working at the intersection of online safety, legal reform, and technology policy. This paper is provided for public interest purposes and does not constitute legal advice. For further information: www.digitaldefence.org