31 May 2021

Review of Model Defamation Provisions
C/o Policy, Reform and Legislation
NSW Department of Communities and Justice
GPO Box 31
Sydney NSW 2001

By email: defamationreview@justice.nsw.gov.au

Dear Chair,

Thank you for the opportunity to provide a written submission as part of the Review of Model Defamation Provisions Stage 2 Discussion Paper.

Twitter is committed to working with the Government, our industry partners, academia, non-government organisations, and wider civil society as we continue to build our shared understanding of the issues and find optimal ways to approach these together.

We support smart regulation, and our focus is on working with governments to ensure that regulation of the digital industry is practical, effective, and feasible to implement while remaining inclusive and keeping core democratic values intact while promoting tech innovation, including Twitter's core commitment to an Open Internet worldwide. In this vein, we support strong mechanisms to protect against defamation and assist in the swift removal of illegal content, while balancing the need to protect principles of free expression to prevent a chilling effect on robust and open public discourse and avoid any unintended harmful consequences.

We trust this written submission will be a useful input to the Department's consultation process. Our written submission also stands together with industry submissions from the Digital Industry Group Inc. (DIGI) and the Communications Alliance. Working with the broader community we will continue to collaborate to create a safe and secure digital ecosystem.

Thank you again for the opportunity to provide input as part of this important legislative reform process.

Kind regards,

[redacted]

Kara Hinesley
Director of Public Policy
Australia and New Zealand

[redacted]

Kathleen Reen
Senior Director of Public Policy
Asia Pacific

Introduction

In response to the Model Defamation Provision ("**MDPs**") Stage 2 Discussion Paper ("**Discussion Paper**"), Twitter's submission will focus on:[1]

1. Definitions and categorisation of Internet intermediaries
2. Liability for publication
3. Power of the courts to order removal of content and reveal the identity of originators
4. Uniform complaints regime for allegedly defamatory content published online
5. Protections and defences

Definitions and categorisation of Internet intermediaries

As currently framed in the Discussion Paper, there are insufficient safeguards in place to protect Internet intermediaries from liability for defamation where they have not authored content. The definitions proposed by the Discussion Paper, which differentiate solely between basic Internet services, digital platforms, and forum administrators, does not contemplate the diversity of Internet intermediaries.

The ACCC Digital Platforms Inquiry definitions of digital platforms are problematic as the categorisation that these platforms are 'considerably more than mere distributors or pure intermediaries of the news content in Australia' is based on an assumed "active" role of all digital platforms in 'selecting, evaluating, ranking and arranging content', and does not take into account the different categories of digital platforms in existence, the distinct ways in which they operate, the variety of ways they are used by people, and the degree of control they have over different types of content featured on the respective services.

Each type of digital service exhibits distinct features in relation to both the potential for defamation and the reasonable steps that can be taken to prevent defamatory content, which necessitates different liability regimes and solutions.

The Discussion Paper describes this distinction between type and conduct in relation to proposed immunity for basic internet services as a 'functions-based approach' or a 'principles-based approach'. The former approach prescribes certain 'functions' (e.g. internet service providers) as attracting immunity, while the latter focuses on features like passivity and neutrality in characterising when and whether an Internet intermediary can be liable for defamation. The common law provides support for this approach.[2]

Content neutrality and algorithm participation

Whether an Internet intermediary is 'content neutral' should be considered as a factor in determining liability for defamatory material for both basic Internet services and digital platforms, rather than solely for basic Internet services.

---

[1] Attorneys-General, 'Review of Model Defamation Provisions – Stage 2' (Discussion Paper, Communities & Justice, NSW Government, April 2021) 47, 3.88 ('Discussion Paper').
[2] *Google Inc v Duffy (2017)* 129 SASR 304 at [206].

The Discussion Paper only describes basic Internet services as 'content neutral'. The corresponding characterisation of digital platforms as 'push[ing] out user-generated content through rankings usually managed by algorithms' is problematic. This is because the distinction between content neutrality and algorithms suggests that these features are opposites and that the promotion of material through algorithms cannot occur by an intermediary on a neutral basis.

The use of algorithms does not prevent a digital platform from still acting as a 'mere conduit' or in a way which is 'content neutral'. Additionally, intermediaries are not intentionally publishing by virtue of an algorithm.

Twitter has invested considerable resources into providing people with a greater level of control and choice over how algorithms affect the content they might see on the service. For example, in the Home Timeline, Twitter has allowed people to choose between viewing the top Tweets first (i.e. the Top Tweets are based on accounts and content the user interacts with the most), or seeing the most recent Tweets first (i.e. Tweets in classic reverse chronological order) since introducing an algorithmic timeline in 2016.[3] This option can be easily identified and executed by tapping the 'sparkle icon' in the main app interface, which we actively communicate in official public communication channels, like company blogs, official accounts, and the Twitter Help Centre.[4]

Further to these public facing mechanisms, we have engaged with rapidly developing, critical subject areas such as algorithmic transparency, choice, and ethics. In April 2021, we announced the establishment of Twitter's Responsible Machine Learning Initiative, led by our internal Machine Learning Ethics, Transparency and Accountability (META) team. This team's responsibilities include driving transparency about our machine learning decisions, how we arrived at them, and vitally, enabling agency and algorithmic choice for people on Twitter[5]. We believe it is important to understand the agency held by the individual when using the Twitter service and the choices available to them regarding how algorithms might affect what they see.

Interaction with the Broadcasting Services Act

Currently under the *Broadcasting Services Act 1992* (Cth) ("**BSA**"), Sch 5, cl 91, certain 'Internet service providers' and 'Internet content hosts' (as defined) are afforded protection from liability. Specifically, Internet service providers and Internet content hosts can rely on clause 91 as a defence where they were not aware of the nature of Internet content and where a service provider or content host would be required to monitor, make inquiries about, or keep records of Internet content.

---

[3] Blog.twitter.com. Never Miss Important Tweets From People You Follow. [online] Available at: https://blog.twitter.com/official/en_us/a/2016/never-miss-important-tweets-from-people-you-follow.html <Accessed May 2021>
[4] Twitter Support, 2018, Available at: https://twitter.com/TwitterSupport/status/1075506036818104320 <Accessed May 2021>; Twitter Help Centre, 2021. Available at: https://help.twitter.com/en/using-twitter/twitter-timeline <Accessed May 2021>.
[5] Blog.twitter.com. Introducing our Responsible Machine Learning Initiative, https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative.html, <Accessed May 2021>

As noted in the Discussion Paper at 3.19, for the purposes of clause 91(1) of the BSA as applied to state defamation laws, it is not clear whether a general complaint to an Internet intermediary is sufficient to make it 'aware of the nature of the internet content' or whether a complaint must specify the defamatory nature of the content. It is also not clear whether a court judgment finding the material in question defamatory is required before the Internet intermediary loses the immunity in clause 91(1). This will be explored further below.

The Discussion Paper also points out at 3.21 that there is uncertainty as to the territorial reach of clause 91(1). In *Fairfax Media Publications; Nationwide News Pty Ltd; Australian News Channel Pty Ltd v Voller* [2020] NSWCA 102 ("***Voller***"), Basten JA (in obiter) considered that the better view of clause 91(1) is that it applies only to those internet content hosts which host content on servers located in Australia.

The MDPs need to consider the interaction of clause 91 with the reforms, including the innocent dissemination defence in section 32 of the *Defamation Act 2005* (NSW), to determine when and how basic Internet services fall within the definitions of 'Internet service providers' and 'Internet content hosts.'

Any amendments to the innocent dissemination defence in section 32 of the *Defamation Act 2005* (NSW) should be reconciled with section 91 of the BSA to ensure there is a consistent approach that media publishers who monitor reader comments are not liable for defamatory comments that they do not endorse or adopt.

Definitions established in the MDPs should also be consistent with definitions contained in the BSA to avoid further confusion or lack of clarity about the protection afforded to certain intermediaries.

Liability for publication

Liability attributed to content originators, not intermediaries

The MDPs need to ensure that Internet intermediaries are not treated the same way as originating publishers for third-party content.

The Discussion Paper recognises that the responsibility of an individual or organisation that creates content in the first place 'is not in question' ("the **originator**").

It is important to recognise that Internet intermediaries are often not in the same position as an originator to assess whether content is defamatory. It is not armed with the background information, evidence, or requisite knowledge in order to make this assessment. An intermediary is also not in a position to defend a potentially defamatory statement as it is not in a position to assess whether any defences may apply to the content.

To reflect the practicalities of dissemination of defamatory matter through Internet infrastructure, Internet intermediaries should not be considered publishers unless and until the intermediary is on notice of the content that has been deemed defamatory by a court of law.

Legislative change to respond to common law deficiencies in protecting Internet intermediaries who are not originators

Under Australian common law, Internet intermediaries can be held to be responsible for publication by reason of having provided a platform for use by content originators, over which they may retain some element of control.[6] The availability of the innocent dissemination defence to Internet intermediaries is uncertain.

As outlined in Part 4, legislative amendments would be required to alter the common law and protect Internet intermediaries from being automatically liable for publication in circumstances where it is more appropriate for sole responsibility to lie with the originator.

Voller is currently the subject of an appeal to the High Court, where it is likely that the Court will provide clarity on the issue of who is a publisher, including whether it is a question of strict liability as set out in *Lee v Wilson* (1934) 51 CLR 276, 287: '[a publisher's] liability depends upon mere communication of the defamatory matter to a third person. The communication may be quite unintentional, and the publisher may be unaware of the defamatory matter.'

Currently, the above unsettled common law principles are unclear. The MDPs should precisely identify whether the definition of publisher is formulated on a conduct-over-type categorisation (i.e. one requiring an element of intention and control and delineating the extent of those limits).

Power of the courts to order removal of content and reveal identity of originator

Twitter opposes amendment to the MDPs that codify the power of courts to order removal of content that has been found to be defamatory by a court, or content that is alleged to be defamatory, prior to a final determination. Twitter's submission in relation to this aspect of the Discussion Paper is underpinned by its belief in the fundamental importance of free speech.

Orders to have online content removed prior to trial

The current test for whether an interim injunction should be granted in defamation proceedings was established in the High Court decision of *Australian Broadcasting Corporation v O'Neill* (2006) 227 CLR 57 ("**O'Neill**").

In *O'Neill,* the court held that the power to order an interlocutory injunction in a defamation proceeding should be approached with 'exceptional caution'. The court must balance the value of free speech in considering whether to grant an interim injunction, and this is a significant consideration. Twitter submits that the common law test in *O'Neill* for whether content should be removed prior to trial (that is, by way of interim injunction) remains appropriate. The common law test fairly balances important considerations of free speech with an individual's right to protect their reputation.

[6] *Webb v Bloch* (1928) 41 CLR 331.

It is unnecessary for amendments to be enacted to the MDPs legislating a statutory test providing a court the power to remove online content prior to trial.

The Discussion Paper contemplates a situation in which an order could be made if a court found in a preliminary hearing that a publication met the serious harm threshold that will be introduced by stage one of the MDP reforms. This would be an inappropriate scenario that would create a chilling effect on free speech.

Situations may develop where a court determines in a preliminary hearing that a publication meets the serious harm threshold, and a court orders that particular allegedly defamatory material must be removed. That proceeding could then proceed to trial and the defendant could be successful by way of a defence. Then the situation would be that the publication has already been removed (due to the plaintiff's success in the serious harm proceeding), but with a successful defence subsequently established to the defamatory material. This raises questions regarding if and how a defendant could be compensated for infringement of their rights and wrongful takedown of their publication, as well as the possible liabilities of relevant intermediaries.

Removing content that has been found by a court to be defamatory

The Discussion Paper considers whether courts should have a clear regime to order Internet intermediaries to take-down or disable access to content that has been found to be defamatory regardless of whether the Internet intermediary would be liable as a defendant in the proceeding, and whether or not it is a party to a proceeding.

In Australia, there is currently no case law where a court has ordered an Internet intermediary, such as Twitter, to remove material that has been determined to be defamatory when that Internet intermediary is not a party to the proceeding. There is currently uncertainty as to whether courts have the power to order non-parties to remove defamatory content.

Twitter submits that it is unnecessary to make amendments to the MDPs to provide power for a court to order a third-party remove material that has been determined to be defamatory. In particular, Twitter opposes the introduction of a similar provision as section 13 of the *Defamation Act 2013* (UK) to the MDPs ("**UK Act**").

The UK legislation currently addresses scenarios in which a court can order the removal of content when judgment has been given for a plaintiff in a defamation action. The legislation is as follows:

Section 13(1) – Defamation Act 2013 (UK)

(1) Where a court gives judgment for the claimant in an action for defamation the court may order—

(a) the operator of a website on which the defamatory statement is posted to remove the statement, or

(b) any person who was not the author, editor or publisher of the defamatory statement to stop distributing, selling or exhibiting material containing the statement.

The UK Act has not been the subject of published case law. The primary concern for Twitter is that the legislation operates as an unnecessary restraint on free speech. For example, circumstances would likely arise in which Twitter is required to determine whether content on its platform conveys similar imputations as those which they have been ordered to remove. If Twitter was then required to remove a substantial amount of content that may not be defamatory, this could have a chilling effect on public discussion.

If the MDPs were amended to provide for a statutory test for when a court may order that online content is removed prior to trial, or a regime for when a court may order that content that has been found to be defamatory is to be removed, a number of safeguards must be enshrined in the MDPs to protect the interests of originators and Internet intermediaries.

First, the right to freedom of expression and free speech must be a primary consideration when determining whether to make an order.

Second, the threshold for overcoming free speech primary consideration should be only in 'exceptional circumstances'.

Third, for orders prior to judgment, the applicant must show an extremely strong *prima facie* case that defamation has occurred and no defences apply.

Fourth, the party to which the prospective order applies (for example, an Internet intermediary) and the originator of the content must be given prior written notice of the relevant application or prospective order, and a statutory right to be heard. For example, the originator of the content may have a strong defence to their allegedly defamatory statement.

Fifth, if an interlocutory order is made ordering removal of allegedly defamatory content, the parameters of any order must be clearly identified. A non-exhaustive list may include: the amount of time the Internet intermediary is required to 'block' content; whether the content is only geo-blocked in a particular jurisdiction; and/or whether global removal of the content is required.

Sixth, the MDPs must clearly identify whether an order could be made if the originator of the content objects to the order or resides outside of Australia.

Seventh, for orders prior to judgment, applicants who are unsuccessful in applications to have content removed prior to trial must be subject to costs consequences that are payable forthwith, to protect freedom of speech.

Eighth, for orders subsequent to content being found to be defamatory, the order must clearly specify the relevant defamatory material that requires removal. An Internet intermediary should not be required to remove content with an equivalent meaning (as this

would put too much onus on the Internet intermediary to consider whether other content conveys certain imputations).

Revealing the identities of originators in relation to a potential defamation action

Twitter's mission is to provide a platform where people have the opportunity to exchange ideas and information, and to express their opinions and beliefs. There is a concern that if preliminary discovery applications become more prevalent in Australia, then it may have a chilling effect on freedom of expression. It may be a disincentive for users to engage in public debate and express their opinions and beliefs if they believe their identity and contact details may be revealed to a complainant in the future without strong safeguards in place.

If there is not effective reform in this area, one of the serious concerns for Internet intermediaries is the potential for such orders to be abused. Specifically, Twitter is concerned the preliminary discovery threshold is so low in Australia that it currently does not sufficiently flesh out whether the prospective complainant's primary aim is actually to unmask an pseudonymous originator with the potential to harass or intimidate that person (e.g. especially if personal details, such as a mobile number, are ordered to be handed over as to a legitimate application). In the United States, the threshold for making such orders is much higher in recognition of the possible chilling effect on freedom of expression.

Currently, a complainant is able to make a preliminary discovery application to ascertain the identity of the originator posting allegedly defamatory material. Twitter would recommend that amendments be made to the MDPs to codify more stringent safeguards when a court may order an Internet intermediary to disclose the identity of a user under such circumstances.

Recent Australian Federal Court decisions in *Kukulka v Google LLC* [2020] FCA 1229 and *Kabbabe v Google LLC* [2020] FCA 126 highlight the ease with which complainants (otherwise known as prospective applicants) have successfully sought provision of information as to the identity of an unknown originator of allegedly defamatory material as part of preliminary discovery against the Internet intermediary.

Rule 7.22 of the *Federal Court Rules 2011* requires that to obtain an order, the complainant must satisfy the Court that:

- *there may be a right for the Prospective Applicant to obtain relief against the prospective respondent;*
- *the Prospective Applicant is unable to ascertain the description of the prospective respondent, notwithstanding reasonable inquiries having been undertaken in the circumstances; and*
- *another person knows, or is likely to know, the description of the prospective respondent, or has, or is likely to have had, control of a document or information that would help ascertain that description.*

The current threshold for granting orders for preliminary discovery to identify a prospective defendant is relatively low in Australia compared with the United Kingdom (UK) and Ontario, Canada, discussed in further detail below. For example, in Australia the complainant is not

required to demonstrate a prima facie cause of action in defamation, nor are they required to demonstrate they are acting in good faith.

As is evident from Rule 7.22 set out above, there is currently no express requirement for the court to have regard to competing considerations such as privacy of users, freedom of expression, or the protection of whistleblowers when considering pre-action discovery applications. This is of great concern to Twitter as there are likely to be issues regarding the conflicts of laws if the originator is not a citizen or resides outside of Australia.

The UK approach

In the UK, orders to 'innocent' third parties to disclose the identity of alleged originators of defamatory content are known as 'Norwich Pharmacal' orders which derived from the UK case of *Norwich Pharmacal v Commissioners of Customs and Excise* [1974] AC 133.

The Court exercises its equitable jurisdiction and therefore under the common law a complainant is required to prove:

a. *a wrong must have been carried out, or arguably carried out, by an prospective defendant;*
b. *there must be the need for an order to enable action to be brought against the prospective defendant; and*
c. *the person/company against whom the order is sought must:*
    i. *be mixed up in so as to have facilitated the wrongdoing; and*
    ii. *be able or likely to be able to provide the information necessary to enable the prospective defendant to be sued.*

Unlike in Australia, UK courts are expressly required to take into account countervailing human rights considerations, such as data rights, rights of privacy, and the right of freedom of expression. Such considerations have set the threshold for the test. Under the first limb of the test, the complainant must show that the prospective defendant 'arguably', or (as a recent case has put it) 'well arguably', committed wrongdoing, and, under the second limb, the making of the order must be a 'necessity', which introduces considerations of alternative mechanisms and of proportionality.

As mentioned above, the current threshold for granting orders for preliminary discovery to identify a prospective defendant in Australia is relatively low compared to other similar jurisdictions. There is no requirement in Australia for a complainant to demonstrate a prima facie, or at least an arguable, cause of action in defamation, as required under the UK approach.

The Ontario approach

In March 2020, the Law Commission of Ontario, Canada, released its final report following its Defamation Law in the Internet Age review process ("**LCO Report**").

According to the LCO Report, Norwich Pharmacal type orders are often directed to Internet intermediaries. The test was established by the Ontario Divisional Court in *Warman v.*

Wilkins-Fournier[7]*.* In this case, the Court held that the Rules of Civil Procedure must be interpreted in a manner consistent with Charter rights and values, including the right of freedom of expression and privacy interests.

The Court established a four-part test for determining whether a third party must disclose the identity of an anonymous online user. The court must consider whether:

a. *the unknown alleged wrongdoer had a reasonable expectation of anonymity;*
b. *the applicant had a prima facie case of defamation and was acting in good faith;*
c. *the applicant had taken reasonable steps to identify the anonymous party and had been unable to do so; and*
d. *the public interest favouring disclosure outweighed the freedom of expression and privacy interests of the unknown alleged wrongdoers.*

The Court also opined that anonymous speech should be afforded some degree of protection as a component of freedom of expression as protection for anonymous speech encourages more speech.[8] The Court also considered that it enhances public discourse, particularly in cases where public interest speech is motivated by fear of persecution or social ostracism. Anonymous speech also allows the author's message to be heard without being coloured by the author's identity and permits sensitive information to be conveyed without embarrassment.

Recommendations with respect to power of courts to order removal of content and reveal identity of originator

Twitter considers the privacy and protection of its users to be of the utmost importance. It considers the application of the Federal Court Rules, and also the similar provisions within the state civil procedure rules, in pre-action discovery application where a complainant is seeking the user's details to be an insufficient protection of user's private information, and a threat to the general public's confidence in an ability to publish sensitive information, opinions, and beliefs without sufficient safeguards in place.

Twitter would be opposed to a regime whereby it would be required to hand over the originator's details except where a prospective complainant has made an application to an Australian court which satisfies a proscribed test.

Twitter is concerned to limit such an order being made to Australian-based users only. It is submitted that Australia should consider adopting a similar test to that established in the *Warman v. Wilkins-Fournier* case. Twitter considers this test to appropriately balance the interests in pseudonymous/anonymous free speech and privacy on the one hand, and reputation and the administration of justice on the other hand.

In relation to the recommendations that Internet intermediaries must retain any records relating to an originating poster for a period of one year, once they have been put on notice of a potential application for an order revealing the identity of the relevant originating poster, there may be issues regarding conflicts of laws if the originating poster is a citizen, resident,

[7] *Warman v. Wilkins-Fournier*, 2010 ONSC 2126.
[8] *Ibid.*

or otherwise resides outside of Australia. We would also seek clarity on the position of the Office of the Australian Information Commissioner (OAIC) regarding any proposal for a new personal or identifying information retention process, as part of the *Privacy Act 1988 (Cth) (Privacy Act)* review that is currently underway.

Uniform complaints regime for allegedly defamatory content published online

We support the notion of Internet intermediaries having clear procedures to respond to requests from complainants through a notice and takedown scheme. However, numerous issues would exist if there was the introduction of a prescriptive complaints process modelled on the UK *Defamation Act*.

The current UK Act model effectively places the onus on the intermediary to remove the content rather than having the first approach remain with the author/originator of the content in question.

This risks hindering freedom of expression, as content is required to be removed by the intermediary to mitigate liability, when that content may not in fact be defamatory of the complainant. The Internet intermediary will very rarely be in a position to determine if the material complained of is defamatory. The risk of non-defamatory material being captured is high, as is the possibility of the complaints mechanism being overused by complainants' who simply do not like content posted. The originator is in the best position to justify any statement complained of, and without the cooperation of the originator, the Internet intermediary is unable to justify or determine the accuracy of the statement.

Clarity is also required as to the way in which a complaints notice will interact with a concerns notice as prescribed in the MDPs. It is not clear whether a complaints notice is a preliminary step to a concerns notice, for the purpose of a complainant accessing contact information of an originator from an Internet intermediary, or if it is lodged at the same time as a concerns notice, or is intended to function as a concerns notice itself.

A strict and punitive complaints regime, such as the UK Act model, would be an inappropriate fit for the MDPs. Instead of prescribing all the ways in which an Internet intermediary can discharge its burden, the burden of establishing the validity of the remedy (i.e. removal) should lie on the complainant. As it stands, the UK Act model:

- discourages free speech by incentivising Internet intermediaries to adopt a 'remove now, ask questions later' approach; and
- improperly gives complainants a cause of action they may not otherwise have had if a decision is made not to remove a post, or if the very strict regulations are not complied with.

By placing the onus on Internet intermediaries, section 5 of the UK Act unfairly penalises Internet intermediaries for a decision not to remove. It not only opens them up to the prospect of protracted and costly litigation (in circumstances where the common law may not currently recognise them as a publisher), it fosters a chilling culture of 'remove first, ask questions later'.

Unlike the UK Act model, the LCO Report recommended a regime which seeks to "place responsibility for online defamation squarely on the shoulders" of content originators.[9] It does this by requiring the Internet intermediaries to act as go-betweens between complainants and originators.

This approach more appropriately recognises the primary point of defamation law: protection of a complainant's reputation and hurt feelings. Instead of requiring an Internet intermediary to assess the validity of the complaint in a vacuum, it affords it an opportunity to try and understand the complainant's grievance, and where the originator refuses to remove the content, the reasons for their non-compliance with the demand.

That being said, the regime recommended in the LCO Report is not perfect. In circumstances where an Internet intermediary cannot identify or pass on a complaints notice to an originator, or where an originator does not respond, the Internet intermediary has a statutory duty to remove the content from its platform. This is unnecessarily reactionary. There are myriad reasons why an originator may not be contactable, or may not respond. Of concern is the turnaround time for the Internet intermediary to act as the go-between.

In circumstances where the originator does not respond in the prescribed time period, the onus for having the complaint removed should shift back onto the complainant. One way this could be achieved is for the MDPs to establish a regime with the features described at paragraph 3(d)(vii). Such relief should only be granted in exceptional circumstances to protect the rights of originators and Internet intermediaries. A complainant should be required to show exceptional circumstances (and is better positioned to do so, given they possess the relevant facts and evidence underpinning their complaint).[10] While there may be access to justice-type issues with the proposal, such issues are a necessary safeguard to free speech and expression and to protect the rights of originators and Internet intermediaries.

As flagged above, any such order should be limited in its effect. It should not require an Internet intermediary to wholesale remove the publication. Rather, it should require an Internet intermediary to remove the publication on a geographically limited or restricted basis. As the tort is only actionable for publications comprehended in Australia, Internet intermediaries should not be required to act on a worldwide basis at the behest of an Australian complainant or court.

A platform such as Twitter – which empowers originators to be in complete control of what is published on their account – does not influence or control the originator. Further, it does not endorse an originator's publication but, rather, staunchly defends an originator's right to free speech. This primacy of an originator's right to free speech is undermined by a complaints procedure that does not allow for a consideration of the merits of the complaint. Prioritising expediency over substance would be a mistake. The MDPs need to very carefully balance a complainant's wish to have content removed by an Internet intermediary with the public interest in continuing to allow Internet intermediaries to facilitate the free flow of ideas, and allow for robust discussions and debates.

---

[9] LCO Report, 81.
[10] *Chappell v TCN Channel Nine* (1988) 14 NSWLR 153.

Protections and defences

The Discussion Paper raises four options to address the liability of Internet intermediaries. In Twitter's view, option 4 is the most appropriate and compelling.

Immunity for Internet intermediaries for user-generated content

In the United States (US), Section 230 of the *Communications Decency Act 1996* ("**CDA Defence**") provides immunity to "interactive computer services" for the publication of third-party content.[11] In the same way the regime recommended by the LCO Report seeks to "place responsibility for online defamation squarely on the shoulders" of content originators (discussed above), the CDA Defence recognises the importance of the free flow of ideas on the Internet.

While US lawmakers are bound by the First Amendment, the lack of such a constitutional protection on free speech in Australia should not be, and is not, an imposition on its legislature. In fact, it is in the interests of justice for the Australian legislature to do what it can to enshrine free speech and protect its citizens from censorship.

Further, the CDA Defence is not absolute. It catches Internet intermediaries who materially contribute to the unlawfulness (i.e. publication of allegedly defamatory material). In this way it adequately balances the rights of complainants against those of passive and neutral Internet intermediaries.

It must be remembered that the CDA Defence is a very broad protection for Internet intermediaries. The focus of the MDPs is only Australia's defamation laws. In this sense, other allegedly unlawful conduct would not be caught by a CDA Defence-type provision of the MDPs.

Ideally, the MDPs would include a CDA Defence-type provision granting Internet intermediaries immunity from liability for third-party user content published on their platform over which they have not had any editorial control or material input.

Fundamentally, this defence would not radically alter the position of the common law. The Discussion Paper notes "[i]t is currently unclear whether a social media platform is considered a publisher under Australian defamation law" (citation omitted). Another way of putting this is there is no authority for the proposition that social media platforms are liable as publishers of defamatory content published by third-party users.

Given the unclear ambit of section 91 of Schedule 5 of the *Broadcasting Services Act 1991* (Cth) ("**BSA Defence**"), the MDPs could adopt parts of the BSA Defence to provide Internet intermediaries with immunity until such a time that they have actual knowledge of their content being defamatory. This means an Internet intermediary would be responsible for the regulation of hosted content only after that content has been found to be defamatory and it is put on notice of that fact.

[11] *Communications Decency Act 1996*, 47 U.S.C. § 230

Such a defence would be complemented by a clear and uniform complaints notice process. Doing so would achieve four important policy objectives:

1. The promotion of free speech by disincentivising frivolous or vexatious complaints;
2. Providing complainants with appropriate means to try and identify the originator when they may otherwise not be able to;
3. Certainty for Internet intermediary defendants whose presence and business is not limited to Australia; and
4. Promoting online innovation and the growth of a strong digital economy.

Arguments posited against a broad immunity such as the one proposed in this submission include:

- Complainants would have insufficient recourse where they cannot identify the originator;
- It is at odds with the traditional approach to secondary publishers; and
- Internet intermediaries have the ability to encourage as well as mitigate the risk of harm to a persons' reputation online.

Each of these arguments should be rejected on the following bases:

- The MDPs could include a robust complaints notice process to allow for complainants to identify originators, and pending non-engagement or a refusal to remove a publication, take steps to obtain an order requiring removal. Further, there is an emerging body of case law grappling with the liability of forum administrators.[12] This presents another possible avenue for complainants;
- Internet intermediaries are not analogous to traditional secondary publishers. The online world is complex and unique, and the sheer scale of information and material published by third-party users on a platform such as Twitter fundamentally distinguishes it from traditional 'subordinate distributors'; and
- As discussed elsewhere in these submissions, whether an Internet intermediary has the ability to encourage or mitigate harm is ultimately a question of its passivity and neutrality.

Finally, a broad immunity would recognise the relationship Internet intermediaries have with their users. When a user creates an account with a social media platform, they sign up to its terms of service. In most cases, these terms include rules and regulations.

For example, Twitter users are required to abide by *The Twitter Rules* which form part of Twitter's Terms of Service.[13] These Terms of Service regulate originators' use of Twitter and seek to ensure all people can participate in public conversation and debate freely and safely. A broad immunity recognises and preserves the contractual relationship already in place between an Internet intermediary (such as Twitter) and an originator.

---

[12] See, eg, *Fairfax Media Publications; Nationwide News Pty Ltd; Australian News Channel Pty Ltd v Voller* [2020] NSWCA 102.
[13] The Twitter Rules. Available at: https://help.twitter.com/en/rules-and-policies/twitter-rules. Accessed May 2021.

While noting Twitter's preference for a broad immunity for Internet intermediaries, Internet intermediaries need certainty regarding what defences are available in response to claims of defamation.

Ideally, a safe harbour protection will be introduced into legislation which exempts Internet intermediaries from liability for defamatory content posted by an originator using an Internet intermediaries technology or platform, in circumstances where that intermediary is not aware of the potentially defamatory nature of the content. This safe harbour provision should apply to Internet intermediaries who did not create the content nor have knowledge that the content is indefensibly defamatory.

An Australian safe harbour regime should *not* be modelled on section 5 of the *Defamation Act 2013* (UK). While the lack of judicial interpretation by UK courts of section 5 limits the ability to consider the operation of the defence, concerns include:

- Extremely short time periods within which intermediaries must undertake tasks which are both time consuming and demand considerable resources. This includes a period of just 48 hours within which an intermediary must review the initial complaint, and either identify irregularities and respond to the complainant, or identify the originator and seek their response. On receiving an originator's response, an intermediary has just 48 hours within which to determine whether the information provided is genuine.
- At a complainant's request, an intermediary must anonymise their complaint before seeking a response from the originator, requiring the originator to respond without being able to determine whether the complainant truly has any legal basis for their complaint.
- In certain circumstances, such as where the intermediary cannot identify the originator within 48 hours, or where the originator does not respond with all required information within 5 days, the Regulations facilitate the removal of content without any consideration of whether the content is defamatory. The requirement that content be removed because of failure to comply with very short timeframes, even where the relevant parties have made efforts to do so, poses concerns for freedom of expression.
- Section 5 affords a complainant an additional period to remedy defects in their complaint, where the intermediary notifies them within 48 hours of receiving the complaint, but does not extend this to originators responding to a complaint. If the originator fails to provide a response that addresses all requirements within the prescribed period, there is no opportunity for them to remedy their response. Rather, the Regulations provide that the matter complained of is to be removed.
- The unworkable nature of the section 5 defence is exemplified by the lack of judicial consideration since its introduction in 2013: many website operators prefer to avoid the impractical procedures set out in the Regulations in respect of the section 5 defence, instead relying on the existing defences, where such defences are required. An Australian defence modelled on the section 5 defence would prove equally unworkable.

If the safe harbour provisions were enacted in the MDPs, Twitter may be encouraged to remove Tweets when put on notice of their allegedly defamatory content in order to ensure it is not held liable for the Tweets. This would have the unintended consequence of chilling free speech.

Further, complainants may be encouraged to make false accusations regarding defamatory content to have content removed, which would mean it would be difficult for Twitter to establish which complaints were bona fide and which were not. This could cause a detrimental impact on Twitter's business activities by requiring Twitter to employ more people in the legal team to consider any potential legal liabilities relating to each individual complaint.

The need to revise the innocent dissemination defence to accurately reflect online forms of publication

The law in Australia is unsettled as to whether Internet intermediaries have the benefit of the defence of innocent dissemination in relation to online publications.

In its present form, the innocent dissemination defence has shortcomings in its application to the Internet and in particular, to social media, due to the immediacy with which content is published online without an editorial process. The burden on the Internet intermediary to establish that it was not aware that the content was defamatory is high.

The availability of the defence to Internet intermediaries requires clarification as to what will constitute knowledge, or constructive knowledge, of the defamatory matter for the availability of the defence to be lost.

The defendant (i.e. the intermediary in this case) has the burden of proving that it did not know, nor ought reasonably to have known, that the content was defamatory and that its lack of knowledge was not due to negligence. Uncertainty remains as to what constitutes knowledge, or constructive knowledge of defamatory third-party content, to an Internet intermediary in different factual scenarios. Accordingly, section 32 of the MDPs does not provide enough protection for Internet intermediaries such as Twitter.

Given this uncertainty, when content is the subject of a complaint, there is a strong incentive for an Internet intermediary, such as Twitter, to promptly remove the matter to retain the ability to rely on the defence, which raises substantial concerns regarding freedom of expression and information in the public interest.

Two options proposed in the Discussion Paper are:

- clarify the innocent dissemination defence in clause 32 of the MDPs to create a default position that digital platforms (such as Twitter) are not primary distributors; or
- introduce a section to the MDPs that applies a presumption that a digital platform (such as Twitter) is a subordinate distributor.

Both of these options would provide more clarity for Twitter regarding its role as a publisher in Australian defamation law. Either option would provide a default position that a digital platform, such as Twitter, will be entitled to an innocent dissemination defence as opposed to the current position whereby Twitter has to prove it is a subordinate distributor.

The approach adopted in some jurisdictions that Internet intermediaries will have constructive knowledge of the defamatory nature of third party content published in search results, or on digital platforms as soon as the intermediary is made aware of its existence, is inappropriate. Such an approach risks stifling freedom of speech and does not appropriately consider the way in which digital technologies work and the immediacy with which publication takes place.

Ideally, the statutory mechanisms will remove the need for Internet intermediaries to rely on the defence of innocent dissemination, as liability will be clarified at the commencement of a cause of action, reducing costs and enhancing certainty. This is because a safe harbour regime will limit the circumstances in which an Internet intermediary, who does not hold the requisite degree of knowledge of publication of the defamatory matter, will be found to have participated in the publication to give rise to liability.

However, if an Internet intermediary loses the protection of a safe harbour defence, as drafted the innocent dissemination defence does not translate to the way in which content is published in a digital age. Rather, it is tailored to a traditional publishing hierarchy in which editorial processes and filters exist, and does not appropriately consider user generated content through Internet intermediaries.

Conclusion

We trust this written submission, together with the industry submission by DIGI and Communications Alliance, will be useful inputs to the Department's work. In summary, we believe the following issues are key areas for further consideration:

- Definitional clarity in the categorisation of Internet intermediaries;
- Attribution of liability to content originators, not intermediaries;
- Support for a clear and uniform complaints notice process; and
- The necessity to protect free speech and open public debate, avoiding unintended harmful consequences or erosion of democratic values.

Twitter is committed to working with the NSW Government, our industry partners, and other stakeholders to ensure that we have a better understanding of the issues at stake and can find the best way to approach this together. Working with the broader community, we will continue to test, to learn, to share, and to improve, so that our platform remains effective and safe for everyone.