Australian Government | eSafetyCommissioner

eSafety Submission to the Social Media (Anti-Trolling) Bill

eSafety Commissioner Submission

January 2022

Introduction

As the world's first government agency dedicated to fostering a safer and more positive online environment, the eSafety Commissioner ('eSafety') welcomes the opportunity to provide feedback on the *Social Media (Anti-Trolling) Bill 2021* ('the draft Bill').

eSafety recently made a separate submission to the Inquiry into Social Media and Online Safety.[1] That submission, and our position statement on anonymity and identity shielding,[2] provide some background which may be helpful to the Attorney-General's Department ('AGD') in considering these issues.

This submission provides more specific comments on the particular issues the draft Bill raises as they relate to eSafety's experience and remit of making online systems and processes safer and remediating complaints from individuals experiencing online abuse. This includes:

- highlighting the risk of public confusion in relation to defamation, trolling and adult cyber abuse;
- identifying some of the limitations of the proposals, including challenges relating to the collection, verification and utility of information about users; and
- noting some of the potential unintended consequences, such as the possible exclusion of some Australians from participation in social media if they are unable or unwilling to provide verified contact details to services.

It suggests there is benefit in considering a more targeted approach to online defamation within the context of broader online safety reforms and welcomes continued discussion with our colleagues at AGD to promote a joined-up approach to online harms across Government.

About the eSafety Commissioner

eSafety is Australia's national independent regulator for online safety. We lead, coordinate, educate and advise on online safety issues and aim to empower all Australians to have safer, more positive online experiences.

When eSafety was formed in July 2015 (as the Children's eSafety Commissioner), one of the agency's main functions was administering a new regulatory scheme in relation to serious child cyberbullying. eSafety also assumed responsibility for the Online Content

[1] eSafety, Submission to the Inquiry into Social Media and Online Safety [PDF], January 2022 https://www.aph.gov.au/DocumentStore.ashx?id=bcab2847-0137-4fcf-8a4d-f59dc0141d24&subId=719461

[2] eSafety, Anonymity and Identity shielding, 22 January 2021, https://www.esafety.gov.au/industry/tech-trends-and-challenges/anonymity

Scheme set out in Schedules 5 and 7 to the *Broadcasting Services Act 1992* (Cth), which was previously administered by the Australian Communications and Media Authority.

Since then, eSafety's functions have broadened to include administration of a civil penalties scheme in relation to image-based abuse ('IBA', sometimes referred to as 'revenge porn'), the power to issue notices to content and hosting services about abhorrent violent material, and a function related to blocking websites providing access to certain terrorist content during an online crisis event.

With the commencement of the *Online Safety Act 2021* (Cth) ('OSA') from 23 January 2022, eSafety will administer another world-first scheme to address cyber abuse material targeting an Australian adult. While eSafety has been informally assisting adults concerned about abusive and harassing material online since 2017, the OSA provides removal powers backed by civil penalties in relation to material intended to cause serious harm.

Beyond the protections built into our authorising legislation to facilitate removal of harmful content, other fundamental elements of eSafety's successful regulatory model are to deliver compassionate citizen service, prevent online harms through awareness and education, and develop and promote initiatives that can achieve proactive and systemic change.

An overview of the Online Safety Act 2021

On 23 January 2022, the new OSA commenced.

The OSA will expand eSafety's regulatory remit and improve the effectiveness, reach and impact of eSafety's work.

In addition to establishing the Adult Cyber Abuse Scheme and extending eSafety's regulatory reach – both across borders and to a broader array of service providers within the digital ecosystem – the OSA also includes several measures that will seek to make platforms and services safer. These include:

- Taking more of a systems-based approach to regulation by promoting and assessing the compliance of services, including social media services, with the new Basic Online Safety Expectations ('BOSE'). These will include expectations of online service providers, including social media services, to take reasonable steps to prevent anonymous accounts from being used for unlawful or harmful material or activity. For example, by having processes that require verification of identity or ownership of accounts.[3]

- Registering codes for eight sections of the online industry, including social media services, which will require service providers to take more proactive steps to prevent and address the harms associated with 'class 1' material (such as child sexual exploitation material or pro-terror content) and 'class 2' material (content unsuitable for children, such as pornography).[4]
- Clarifying and enhancing eSafety's powers to obtain information in support of regulatory investigations, including about end-users of social media and other services. By opening key lines of inquiry for eSafety to pursue through its regulatory investigations, these powers will enable us to hold individuals accountable for perpetrating harm, including those operating publicly anonymous or pseudonymous accounts.

Adult cyber abuse material

Under the OSA, adult cyber abuse ('ACA') material is made out when there is material that is posted with the apparent intention of causing serious harm, and the material is offensive, harassing or menacing in all the circumstances.[5] 'Serious harm' is to be understood as encompassing serious physical harm, serious psychological harm or serious distress, but does not include 'mere' ordinary emotional reactions such as simple distress, grief, fear or anger.[6]

The test is set deliberately high to recognise the greater resilience and responsibility shouldered by adults for managing their own online experiences, and to limit intrusions into freedom of expression. More information about how the eSafety Commissioner will manage the ACA Scheme can be found in eSafety's related regulatory guidance.[7]

The ACA Scheme is designed as a safety net to ensure the prompt removal of material which is intended to cause serious harm. As set out in the Explanatory Memorandum to the OSA, the definition "is not intended to capture 'reputational harm' caused by

[3] *Online Safety (Basic Online Safety Expectations) Determination 2021*, section 9.

[4] eSafety, Development of industry codes under the Online Safety Act Position Paper, September 2021 https://www.esafety.gov.au/about-us/consultation-cooperation/industry-codes-position-paper#:~:text=The%20new%20Online%20Safety%20Act,eSafety%20to%20register%20the%20codes.

[5] *Online Safety Act 2021* (Cth), section 7.

[6] *Online Safety Act 2021* (Cth), section 5.

[7] eSafety, Adult Cyber Abuse Scheme Regulatory Guidance, December 2021, https://www.esafety.gov.au/about-us/who-we-are/regulatory-schemes#adult-cyber-abuse-scheme

defamatory material”.[8] Nor are eSafety’s investigative processes intended or equipped to arbitrate the truth or falsity of statements made online. However, in some cases, material which may be defamatory could also meet the threshold for ACA if eSafety is satisfied there was an intention to cause serious physical or psychological harm and the material is menacing, harassing or offensive.

ACA material can manifest on websites (such as forums) and messaging services; however, eSafety has predominantly received reports about material posted on social media services. Between 1 July and 31 December 2021, eSafety provided informal assistance to 816 adults in relation to abuse or harassing material online. Reports are on the rise, increasing by more than 50% during financial year 2020-21 compared with financial year 2019-20.

To date, some of the most serious forms of abuse reported to us involve:

- impersonation accounts posting abusive messages to victims’ contacts,
- technology-facilitated stalking (trackers, hacking, surveillance, keyboard monitoring) and
- ‘volumetric’ attacks (also known as ‘pile-ons’ or ‘brigades’) involving many people targeting a single person, often across multiple platforms.

Adult cyber abuse reports also include blackmail, threats and doxing, to intimidation, posting offensive or upsetting imagery and defamation. Whether or not these reports meet the legislative threshold of ACA is to be determined on a case by case basis.

eSafety’s research and reporting trends show certain individuals and groups are disproportionately at risk of online harm or face additional barriers to protecting themselves from harm or accessing support. At-risk groups include women, who make up two-thirds of all reports to eSafety, as well as:

- older people
- people with disability
- Aboriginal and Torres Strait Islander peoples
- people from culturally and linguistically diverse communities, and
- people who identify as LGBTQI+.

[8] Explanatory Memorandum to the *Online Safety Bill* 2021, at p 70: https://parlinfo.aph.gov.au/parlInfo/download/legislation/ems/r6680_ems_3499aa77-c5e0-451e-9b1f-01339b8ad871/upload_pdf/JC001336%20Clean4.pdf;fileType=application%2Fpdf.

We recognise that many intersecting factors influence risk levels and individual experiences of online harm. Accordingly, we shape and prioritise our programs and resources to support, protect and build the capacity of those who are most at risk.[9]

Information-gathering powers

Under the OSA, the eSafety Commissioner will exercise clarified and strengthened powers to gather information. Under Part 13 of the OSA, this includes information about the identity of end-users, and contact details relevant to end-users. The power can be exercised by the eSafety Commissioner in relation to end-users of social media, designated internet and relevant electronic services. Failure to comply with a notice to produce such information attracts a civil penalty.

- **Social Media Services** include social networking platforms such as Facebook, Snapchat and Instagram.[10]
- **Designated Internet Services** allow users to access or download material using an internet carriage service. Examples of the DIS category include websites, image-hosts and file-hosts (or ‘cyberlockers’).
- **Relevant Electronic Services** include instant messaging services and multiplayer gaming services.

The eSafety Commissioner currently seeks ‘basic subscriber information’ (BSI) from digital services using a power to obtain documents or other information via section 173 of the *Broadcasting Services Act 1992* (Cth). The nature, extent and quality of BSI collected and stored by social media services varies widely. While some platforms require users to create profiles using their ‘authentic name’ or the name they go by in everyday life, others provide wide latitude and explicitly enable the use of pseudonymity.[11]

Generally, users are required to provide some or all of the following details at sign-up:

- Name (or pseudonymous username),
- Email address,

[9] eSafety, Protecting voices at risk online, 2021, https://www.esafety.gov.au/diverse-groups/protecting-voices-risk-online.
[10] A social media service is defined via section 13 of the OSA as being an electronic service that enables as its sole or primary purpose online social interaction between 2 or more end-users; that allows end-users to link to, or interact with, some or all of the other end-users; that allows end-users to post material to the service; and any other conditions (if any) as set out in the legislative rules.
[11] The many benefits that a degree of anonymity or pseudonymity may offer are set out in our submission to the inquiry into social media and online safety at page 55.

- Phone number, and/or
- Date of birth.

Many services (such as Facebook, Twitter and TikTok) only require an email address *or* a phone number, not both.[12] Physical address details are generally not collected by platforms.

While users may be asked to confirm these details at sign-up, services do not necessarily ask users to re-verify on a routine basis thereafter, so contact details may become outdated or obsolete. Services also have no way of confirming that an email address refers to a user's primary account.

Some platforms (such as Twitter) require account holders who have breached their terms of service to verify ownership with a phone number or email address.[13] These retained identity details may help eSafety identify and take action against violators who are operating multiple accounts for abusive purposes.

Main challenges presented by the draft Bill

Challenges with basic subscriber information

While digital services may possess some information provided by an end-user during the sign-up process for a new account, billions of user accounts have been established through minimal collection of identity data, and the quality and reliability of BSI that social media services hold can vary. Accordingly, accessing the conditional defence to defamation under the draft Bill is likely to require a significant expansion of services' current practices for the collection and verification of Australian users' information.

It is not presently clear to us that industry has the capability to collect and hold large amounts of personal information in a private and secure manner.[14] We therefore encourage AGD to consider how the introduction of this Bill may impact services' data

[12] Whether both are required can depend on the kind of device and nature of network used to access the sign-up process. See: Facebook, How do I create a Facebook account? https://www.facebook.com/help/mobile-basic/188157731232424; Twitter, Signing up with Twitter, https://help.twitter.com/en/using-twitter/create-twitter-account; Business Insider Australia, How to make a new account on T kTok in 3 different ways, 22 April 2020, https://www.businessinsider.com.au/how-to-make-a-new-account-on-t ktok-2020-4?r=US&IR=T.

[13] Twitter, Our range of enforcement options, https://help.twitter.com/en/rules-and-policies/enforcement-options.

[14] We draw your attention to page 63 of our submission to the inquiry into social media and online safety which sets out initial stakeholder views canvassed through our consultations for the development of an age verification roadmap.

practices and how these impacts may relate to the objectives and outcomes of the privacy law reform also underway.

eSafety also emphasises that, as a result of the inconsistency between online services in relation to how they obtain, validate, store and log BSI, these details do not always have investigative value. While the supporting materials to the draft Bill note the intention to 'unmask' Australian originators of anonymous defamatory comments made on social media,[15] in our experience, obtaining BSI is only a first step in opening further lines of inquiry to try to identify an end-user.

It is not always the case that a person's identity or even contact details can be adequately established through information held by a digital service. For example, while an Australian mobile phone number may be useful information (as explained below) the identity of the subscriber can generally only be ascertained by querying restricted databases or seeking further end-user information from the service provider. Similarly, information relevant to the creation of an email account will generally be held by third parties, such as Google in relation to a Gmail account or a carrier or internet service provider ('ISP') in relation to an email account associated with a broadband home internet subscription.

While eSafety has powers enabling us to follow these lines of inquiry, this type of account information may provide limited value to individuals or entities who seek to effect service of notice relating to legal proceedings but lack online investigative capacity.

More detailed explanations of these and related challenges follow.

Challenges with email addresses

Some social media services prohibit the use of temporary and 'throwaway' email services; however, this requirement is not enforced across the industry. Users can easily create an unattributable email address using any of scores of webmail providers,[16] and then provide that address for the purpose of account creation.

Identifying a person from such an email address will, in most circumstances, be impossible, and there is no guarantee that email accounts are monitored by end-users.

[15] Attorney-General's Department, Social Media (Anti-Trolling) Bill, https://www.ag.gov.au/legal-system/social-media-anti-trolling-bill.

[16] For an explanation of temporary email addresses, see Digital Information World, 10 email services that offer temporary addresses: https://www.digitalinformationworld.com/2020/04/disposable-email-services-that-offer-temporary-addresses.html.

Challenges with mobile phones

In Australia, it is more difficult to obtain a 'burner' mobile phone number than in other jurisdictions.[17] The Australian Communications and Media Authority ('ACMA') has established rules[18] which require telecommunications companies to obtain certain information from customers and verify their identity before activating a prepaid mobile service.[19]

The information obtained includes the customer's name, home address and in most instances, date of birth. The regulatory arrangements set out the methods that telecommunications companies can use to verify identity, including:

- a government online verification service which undertakes a blind check to match the information on a government issued document (e.g., a Medicare Card) against the information held by the issuer of the document
- a visual identity check which requires a higher standard of verification if the customer has more than 5 activated services, and
- an existing prepaid or post-paid account with the telecommunications company.

Authorised agencies (not private individuals) can query name and address data linked to a person's mobile phone service via the Integrated Public Number Database ('IPND') regulated by the ACMA. Given the relatively high quality of data verification processes related to registration of mobile services, address details are likely quite accurate when a customer first activates a service, though there is little incentive for mobile customers to update their address details with their provider if they subsequently move house, and of course any system can be 'gamed', for example, through identity theft.

Basic subscriber information: IP address logging

BSI can also include IP addresses (a unique address that identifies a device on the internet or a local network). These may be recorded in log entries created when a user accesses a social media platform. The IP address recorded will vary depending on the network used to access the internet. A person accessing the internet from a home Wi-

[17] For an explanation of burner phones, see How-To Geek, What is Burner Phone, and When Should You Use One?, 17 February 2021 https://www.howtogeek.com/712588/what-is-a-burner-phone-and-when-should-you-use-one/.

[18] The *Telecommunications (Service Provider — Identity Checks for Prepaid Mobile Carriage Services) Determination 2017.*

[19] The information telecommunications companies obtain varies depending on whether the customer is a purchaser or service activator. Most telecommunications companies have systems and processes in place to obtain information from, and verify the identity of, a service activator.

Fi network will present a different IP address to the address used when accessing the internet using their mobile network.

Generally, the IP address that is logged will be an IPv4 address. These addresses are familiar to most internet users, looking like this: 103.29.195.64 – the IP address associated with esafety.gov.au. Every device must be assigned an IP address when connecting to the internet. The IPv4 address space provides for about 4.3 billion unique IP addresses. With the growth of the internet (especially the Internet of Things) this address space was exhausted around 2011. That is, there are now more devices connected to the internet than there are unique IP addresses to assign.

When connecting to the internet using a mobile device, it is common for a carrier or ISP to assign one IP address to many devices at once. The individual sessions for each of these devices are differentiated from one another (both for the purpose of connectivity to remote hosts and billing) by being assigned a unique source port through a technology known as Port Address Translation ('PAT'). This combination of IP address with source port should be capable of uniquely identifying a specific user session. If known, the PAT mapping between IP address and source port can be provided to a carrier (such as Telstra) to identify the relevant subscriber.

Challenges with IP addresses

However, it is not always possible to obtain fully-mapped IP addresses/source ports from social media services. Sometimes, the IP address will be logged but the source port will not. Whether both data-points are preserved can depend on network and administrative settings applied by the platforms, but there may be other technical reasons limiting the ability of platforms to supply both when required.

This challenge may be alleviated through IPv6. Given the exhaustion of the IPv4 address space, many network operators are now enabling the successor protocol to IPv4, IPv6 (which is expressed using hexadecimal notation, e.g., Google's IPv6 address is 2607:f8b0:4023:1002::71).

The IPv6 addressing protocol has a far larger addressing space than IPv4 (128-bit vs 32-bit), providing for much more reliable and efficient allocation of IP addresses to internet-connected devices. In theory, this means that where a carrier has implemented IPv6 it is likely that a unique IPv6 address will be assigned to a specific subscriber (rather than relying on a combination of IPv4 and source port). If this is the case, then an IPv6 address may be enough for a carrier to identify a subscriber.

Whether IPv4 and/or IPv6 data is available to be captured and logged is beyond the control of a social media service. The decision about implementing one or both is taken by a carrier or ISP, and only about 30% of internet traffic has an IPv6 address allocated. Despite the improved security and routing efficiencies inherent to IPv6, adoption does come with a requirement to upgrade infrastructure – IPv4 routers are not compatible with IPv6, requiring so-called 'dual-stack' procurement. Not all carriers and ISPs will be able to afford transition to IPv6, and the majority of end-user devices are still only compatible with IPv4. Some estimates note that IPv6 will not be fully adopted until the middle of next century. The pressure to address the shortfall in the IPv6 address space has been significantly ameliorated through introduction of PAT.

Challenges with reliance on geolocation data to show location of commenter

Under the draft Bill, a comment is made in Australia if an end-user of a social media service has posted a comment on a page of the service, and the end-user appears to have been located in Australia when they posted the comment. Establishing the apparent location relies on 'geolocation technology deployed by the provider of the service'.

eSafety suggests that reliance on the service's geolocation technology may limit the efficacy of the draft Bill. This is because geolocation will often depend in large part on mapping IP addresses to their related networks and establishing rough geographic contiguity. This may not always provide an accurate indication of a user's jurisdiction, although the risk of mis-identification of rough location within the Australian context is low given the isolated nature of domestic internet infrastructure.

A more likely challenge arises when one considers the impact of IP obfuscation technologies on the capacity for services to achieve accurate geolocation.[20] For example, the use of Virtual Private Networks (VPNs) and proxy servers make resolution of an end-user's true IP address near impossible. A VPN is a service, usually paid, which encrypts the connection between an end-user's device and the VPN server. In addition to encrypting internet traffic, VPNs also often allow an end-user to select a geographical location from which a connection to the destination website or service will be made.

Proxies permit users to route their internet traffic through one or more third-party servers around the world. Many proxies are free to use and allow end-users to chain one or more to act as routing nodes for internet traffic. This makes it impossible for a social media service to accurately gauge an end-user's geographical location if IP address geolocation comprises a component of its geolocation technology. Of course, a social media service may use other data to establish a user's jurisdiction. These could include analysis of user behavioural signals, user content analysis, and other signals such as Wi-Fi positioning.

[20] It is important to note that there are legitimate reasons for people to employ these types of tools, as outlined in our submission to the inquiry into social media and online safety.

Focus on social media services

The draft Bill focuses on defamatory comments posted to social media services. The definition of 'social media service' relies on the definition contained within the OSA.[21] Otherwise, an electronic service can be specified as a social media service in the legislative rules.[22]

By limiting the draft Bill's scope to social media services, some defamatory material that is customarily shared via other platforms may be missed. For example, material posted in groups on popular encrypted messaging services such as WhatsApp and Telegram would likely not be caught by the draft Bill. This is because their sole or primary purpose is not to enable online social interaction, and there is more limited functionality in relation to the posting of material. Given the size of the Australian user-base of these services (WhatsApp has close to 6 million Australian users)[23] and the corresponding likelihood that they are or will be a vector for potentially defamatory commentary, their inclusion in the draft Bill as relevant electronic services may be useful.

Potential unintended consequences of the Bill

eSafety has identified a number of potential unintended consequences which may follow legislating the draft Bill as written in its present form.

Risk of confusion about the Bill's objectives

The first is that there is a risk of public confusion over what the Bill seeks to achieve if its present name is preserved. eSafety provides guidance on our website about what type of behaviour is generally considered to be 'trolling', noting that:

> *'trolling is when someone makes a deliberatively provocative comment or post and waits for people to take the bait … Trolling is not when someone makes a personal attack'.*[24]

[21] Draft Social Media (Anti-Trolling) Bill (Social Media Bill), section 6.
[22] Draft Social Media (Anti-Trolling) Bill (Social Media Bill), section 6.
[23] Media Week, Social Media Statistics August 2021: Facebook narrowly ahead of YouTube, 3 September 2021 https://www.mediaweek.com.au/social-media-statistics-august-2021-facebook-narrowly-ahead-of-youtube/
[24] eSafety, Trolling https://www.esafety.gov.au/young-people/trolling

Trolling has become a common place term to describe a wide array of abuse. Over time this could serve to normalise or trivialise the very real suffering that those seeking to deliberately menace, or abuse others inflict on their targets.

With the ACA Scheme commencing on 23 January, eSafety has developed public messaging and regulatory guidance to clarify the definition of ACA and manage expectations about what the scheme will and will not cover. In the reports that eSafety has received to date, we have observed a tendency to conflate ACA with defamation and have released information to explain the difference.[25]

We believe there is a risk that a law which seeks to address defamation but uses the terminology of trolling may compound the confusion about where ACA stops and defamation begins, and where to go for help with each of these issues.

Difficulty in verifying relevant contact details for Australian user base

The second potential unintended consequence relates to the structure of the draft Bill's complaints scheme. For a provider of a social media service to avoid being held liable for defamation, the Bill proposes a conditional defence requiring that the provider has a complaints scheme in place. This complaints scheme must satisfy the requirement of being able, should a commenter consent, to furnish upon a complainant the commenter's relevant contact details.[26] These are defined as including the commenter's name, their email address, and a phone number (in addition to any other requirements set out in the legislative rules).[27] Compliance with the scheme requires a provider of a service to be able to provide relevant contact details of a commenter located in Australia.

The draft Bill does not make allowance for the requirements to be met prospectively from date of commencement. As a consequence, it appears to eSafety that social media service providers would be incentivised to take steps to retroactively collect and validate the relevant contact details for *all* of their existing Australian users, in case those users end up engaging in defamation in the future, if they wish to access the conditional defence from defamation liability. This would be a complicated and arguably burdensome requirement for industry, with Facebook alone counting 17 million Australian monthly users.[28] eSafety believes some users may be hesitant to provide

[25] eSafety, What's the difference between serious online abuse and defamation?, 11 October 2021, https://www.esafety.gov.au/newsroom/blogs/difference-between-serious-online-abuse-and-defamation

[26] Draft Social Media (Anti-Trolling) Bill (Social Media Bill), section 15.

[27] Draft Social Media (Anti-Trolling) Bill (Social Media Bill), section 15.

[28] Social Media News, Social Media Statistics Australia – November 2021, 1 December 2021. https://www.socialmedianews.com.au/social-media-statistics-australia-september-2021.

these details to services due to lack of trust, and others may have difficulty complying, as set out in the next section.

As noted by the eSafety Commissioner during October 2021 Senate Estimates, ‘I think there are incremental steps we could make. Getting rid of anonymity or even [the use of] pseudonyms on the internet is going to be a very hard thing to achieve.’[29] eSafety’s submission to the Inquiry into Social Media and Online Safety[30] reiterates this point, noting the many potential safety and privacy benefits that a degree of anonymity or pseudonymity can provide to users, particularly those who may be at risk of oppression, stigma or abuse. The submission also emphasises that online safety interventions should be targeted, proportionate and effective to resolve clearly identified problems, with the rights and best interests of users a paramount consideration.

Potential for low-income households to be disadvantaged

eSafety notes the potential for the compliance steps outlined above to effectively limit access to social media for some Australian users. If a person is unwilling or unable to furnish both an Australian mobile phone number and email address, they may be denied access to widely used services such as Facebook if those services determine that the cost of a blanket exclusion for unverified users is lower than the cost of their potential defamation liability. The practical effect may be to miss out on opportunities to stay connected with family and friends, keep up to date with essential information (such as public health messaging), and find support online (for example, support services offered by Kids Helpline and Lifeline).

This limit to access would likely disproportionately affect lower-income households. Research published during 2020 by the NSW Council of Social Service found that 98% of those in low and low-middle income households possess a mobile phone. However, 68% of those experiencing the greatest housing stress reported being unable to use mobile data because of either exceeding their data allowance or running out of credits.[31] For these users of social media services, an attempt by a provider to validate a mobile service to comply with the Bill’s complaint scheme may fail if the user cannot access data. Such a user may be excluded from access as a result, leading to inequitable outcomes across the Australian end-user base.

[29] Senate Estimates Hansard, Environment and Communications Legislations Committee, 21 October 2021. https://www.aph.gov.au/Parliamentary_Business/Hansard/Hansard_Display?bid=committees/estimate/efe5d2c5-9316-4170-adc7-f1aafe99ca9e/&sid=0006.

[30] eSafety, Submission to the Inquiry into Social Media and Online Safety [PDF], January 2022. https://www.aph.gov.au/DocumentStore.ashx?id=bcab2847-0137-4fcf-8a4d-f59dc0141d24&subId=719461

[31] NSW Council of Social Service, *The Cost of Staying Connected*, 2019 https://www.ncoss.org.au/wp-content/uploads/2020/04/AB-Telecommunications-CoLiN-results-final.pdf.

Challenges with complaint scheme threshold issues

We note that the threshold issue under the complaints scheme requires self-assessment from a complainant. Specifically, in order for a complainant to make a complaint to a provider of a social media service, the complainant must have reason to believe that there may be a right for the complainant to obtain relief against an end-user of the service (the commenter) in a defamation proceeding that relates to a comment posted on a page by the commenter.

This requires the complainant to have at least a working knowledge of the law of defamation, and be able to distinguish between offensive, hurtful and insulting comments on the one hand, and material harmful to a person's reputation on the other. In our experience, the general degree of awareness of what constitutes potentially defamatory material posted online is likely low among most Australian end-users of social media services. The need to self-assess whether there is a right to obtain relief under the law of defamation is likely to lead to many complaints under the scheme falling short of what might be considered actionable under the law of defamation. The provision of free legal information and advice could assist, as highlighted below.

Risks to privacy and safety

eSafety's welcome the draft Bill's provision that a court may refuse to make an order requiring the disclosure of a commenter's contact details if satisfied that this is likely to present a risk to the commenter's safety. However, we query how a court is to obtain the factual information it would need to make this assessment – particularly if the commenter has not been involved in the proceedings.

In most cases, privacy and confidentiality considerations prevent eSafety from sharing any BSI we may obtain through the course of an investigation with others, including the relevant complainant. We believe this is an important protection against the risk of retaliation.

While BSI from social media services on its own is often insufficient to identify a specific person with certainty (as set out above), this data can be used for harmful purposes, including doxing.[32] For example, a person could post an email address or phone number online and invite others to dole out punishment. eSafety's research into adults' negative online experiences shows that around 95% of adults who reported behaving negatively online to someone in the 12 months to August 2019 had also been

[32] eSafety, Doxing trends and challenges – position statement, 29 May 2020, https://www.esafety.gov.au/industry/tech-trends-and-challenges/doxing.

the target of a negative online experience during the same period, demonstrating that harmful online interactions between adults are rarely straightforward.[33]

Areas for consideration

Online services should take proactive and reasonable steps to deter misuse of their platforms, empower victims to stay safe when problems do occur, and hold perpetrators accountable where online abuse is deemed serious and/or ongoing. eSafety believes that services can improve practices across each of these three areas without necessarily having to gather large amounts of identifying information about all their users. These improvements should aim to remediate harms that individuals are experiencing, as well as to prevent and address these harms on a systemic basis.

Bolstering the proposed complaints scheme

eSafety think it is unlikely that a commenter responsible for material that is potentially capable of leading to commencement of defamation proceedings will consent to being exposed as a defendant in litigation. However, we do see value in the proposed complaints scheme. We believe the conditional defence to defamation for services can be linked to this scheme in a way that balances the imperatives of user safety, privacy and security.

We suggest that, at the same time the service seeks consent to share the commenter's details, the service could:

- Ask the commenter if they have any safety concerns related to the disclosure of their information, to ensure this information is available to inform subsequent proceedings;
- Preserve any account details the service already has about the commenter, to mitigate against the risk that relevant details might be deleted;
- Take reasonable steps to collect, verify and retain contact details such as a phone number for the commenter (rather than doing so for all Australian users of the service); and
- Indicate to the complainant whether the service has obtained any BSI about the commenter, whether the commenter has consented to share it and whether the

[33] eSafety, Adult's negative online experiences, August 2020, https://www.esafety.gov.au/research/adults-negative-online-experiences.

commenter has raised other concerns as to why it should not be shared (without going into any detail regarding the specific concerns).

Assuming the commenter withholds consent to disclose their details, this process nevertheless provides benefits to both the commenter and the complainant. It places the commenter (and the service) on notice that the material may be defamatory and provides an opportunity to consider removing it or to provide the service with additional context about the situation, including any safety concerns. It also provides the complainant with sufficient information to determine whether there may be value in pursuing an application for an end-user information disclosure order.

We suggest the conditional defence to defamation could be available to a service which:

- Provides the information in response to such an order (noting the likelihood that this will not 'unmask' the commenter and further investigation will be needed);
- Can demonstrate that it took reasonable steps to collect, verify and preserve the information but was unable to do so; or
- Can demonstrate that the commenter has raised legitimate safety concerns about why it should not be shared.

We believe this is a more targeted and privacy-preserving response to individual defamation matters, as it does not require the collection and regular verification of all Australian users' information in case they engage in defamation in the future.

Systemic issues, such as greater consistency in BSI practices and the prevention of creation of new accounts for the purpose of continuing to defame, harass or abuse a person, can be addressed through Safety by Design and the Basic Online Safety Expectations.

Safety by Design

Protecting and safeguarding users online is a global concern. To reduce risks and counter threats, we need a proactive and systemic approach that addresses the complex societal situations and behaviours that manifest in the online world. Online risks and harms are often inter-related and inter-connected and occur across the entire spectrum of online devices, services and platforms.

Advancements in technology, machine-learning and artificial intelligence have the potential to radically transform user experiences and safety online. At eSafety, we recognise the importance of proactively and consciously considering user safety as a standard risk mitigation and development process, rather than retrofitting safety considerations after online harms emerge or damage has occurred.

There is global recognition of the need to create and develop responsible technology, that takes a holistic view of consumer welfare and considers the broader societal impacts of online and technological products and services. However, a patchwork of legislation, regulation and governance structures contribute to inconsistent and fragmented systems, resulting in gaps in oversight and accountability.

eSafety's Safety by Design initiative helps to overcome some of these challenges. Safety by Design guides and supports companies to assess risks up front, encouraging informed, transparent and accountable decision-making process that work to minimise harm whilst building user safety into the design, development and deployment of online products and services.[34] From our experience of regulating for specific online harms within Australia, we concluded that the only way to truly get ahead of the multiple safety challenges online is through a combination of cultural and behavioural change, and the development of a safer online infrastructure.

Three areas that we believe online services should improve in order mitigate against online harms, including but not limited to defamation, are:

- **Deterring misuse** – Preventing poor behaviour is critical. Services should continue to innovate ways to reduce the sense of disinhibition online, to promote pro-social norms on their platforms and to create friction to make it more difficult to abuse others. An example of this is 'nudge' technology which sends an educative prompt, warning or reminder to users before they send or post a comment or content which is likely to be harmful, asking them if they are sure they wish to continue. On some platforms such prompts also direct users to the platform terms of service and or third party support services.

- **Empowering victims** – When problems do occur, users should have easy access to appropriate tools and information to keep themselves safe, and to seek swift and effective help from the services where the harm is happening. Blocking, muting and reporting mechanisms are basic examples of these types of 'conversation controls' and are most effective when accessible in platform.

- **Holding perpetrators accountable** – Where online abuse is serious and/or ongoing, there must be avenues to hold perpetrators accountable. This requires effective enforcement of consequences for terms of use violations, such as temporary account suspensions and permanent bans. The ability to create multiple accounts to abuse a victim is an issue we have seen in our investigations, where new accounts have been provided to existing users who continue to bully or threaten a victim after they successfully block and report the initial offending account. Rather than requiring every user to verify their identity upon the creation of an account, services could improve their processes for detecting new accounts likely engage in abusive behaviour through technological

[34] eSafety, Safety By Design, https://www.esafety.gov.au/industry/safety-by-design.

means such as device or offending IP address detection. Where potential abuse is identified, services could require the account holder to verify certain identity or contact details to re-gain access to their suspended account(s) or to facilitate investigations through appropriate legal processes.

Through a combination of these measures, services can achieve a balance between safety, privacy and security for their users.

Basic Online Safety Expectations

In addition to Safety by Design, the OSA provides eSafety with several options to drive improvements across deterrence, empowerment and accountability. Most notably, the Basic Online Safety Expectations (BOSE) will set a higher benchmark for the steps we expect services to take to keep people safe.

The Minister for Communications, Urban Infrastructure, Cities and the Arts has determined the BOSE through the Online Safety (Basic Online Safety Expectations) Determination 2022 (Determination), registered on 23 January 2022.[35]

The Determination includes a specific expectation regarding anonymous accounts, placing the onus on services to take reasonable steps to prevent those accounts from being used for harmful or unlawful material or activity. The Determination provides examples of reasonable steps services could implement, including having processes that prevent repeated use of anonymous accounts to cause harm as well as processes that require verification of identity or ownership of accounts.

eSafety will have the power to require services to report on the steps they are taking to meet the BOSE, and to explain how they are adequate and effective to prevent and address online harm. This will enable us to address online harms – including those exacerbated by identity shielding – on a systemic basis.

The obligation to respond to a reporting requirement is enforceable and backed by civil penalties and other enforcement mechanisms. eSafety can also publish statements about the extent to which services are meeting the expectations. We believe this has great potential to help improve safety standards, and to bring greater accountability to services whose transparency to date has been highly selective and uneven.

[35] https://www.legislation.gov.au/Details/F2022L00062

While we appreciate that the focus of the draft Bill is online defamation, we note that policy consideration of anonymity, pseudonymity and identity verification/shielding should be joined-up across the full range of online harms. Where AGD and other stakeholders identify instances where it appears that services are not taking reasonable steps to collect and verify BSI or other data (such as device identifiers) in order to enforce consequences where there appear to be breaches of terms of service, we believe the BOSE can contribute to driving improvements in services' practices.

Access to justice issues

The difficulty in confirming the identity and contact details of the relevant account holder is one barrier to resolving defamation matters.

However, there are also broader issues in relation to access to justice. Many people cannot afford legal advice, as free community legal centres tend to prioritise other issues due to limited resources.

One option to address this broader issue would be for AGD to develop, or fund the creation of, online information about defamation. This could be supported by the funding of community legal centres to provide advice and assistance to send defamation concerns notices on behalf of members of the community who are subject to online defamation. While such a program would not necessarily be able to fund litigation, we know from our own experience that notices can be helpful in their own right without escalating to enforcement in court.

We believe the availability of legal services able to advise about online defamation would be highly valuable to support our new ACA Scheme by enabling us to make referrals where we receive complaints involving potential defamation. It would allow us to continue to provide a critical service where we triage online abuse matters and refer them on seamlessly to those best equipped to help.

Conclusion

eSafety appreciates the opportunity to provide feedback on the draft Bill.

eSafety believes our holistic functions and powers across the pillars of prevention, protection and proactive and systemic change – strengthened by the OSA – will continue to give Australians safer, more positive experiences online by placing greater onus on services across the online ecosystem to prevent and address both individual and systemic harms.

While issues such as defamation fall outside the scope of our remit, we are keen to continue working with AGD and others to promote targeted and consistent approaches to combat the full array of online harms, and to ensure information and services are available to support ACA complainants whose matters may involve defamation or other legal claims.

***