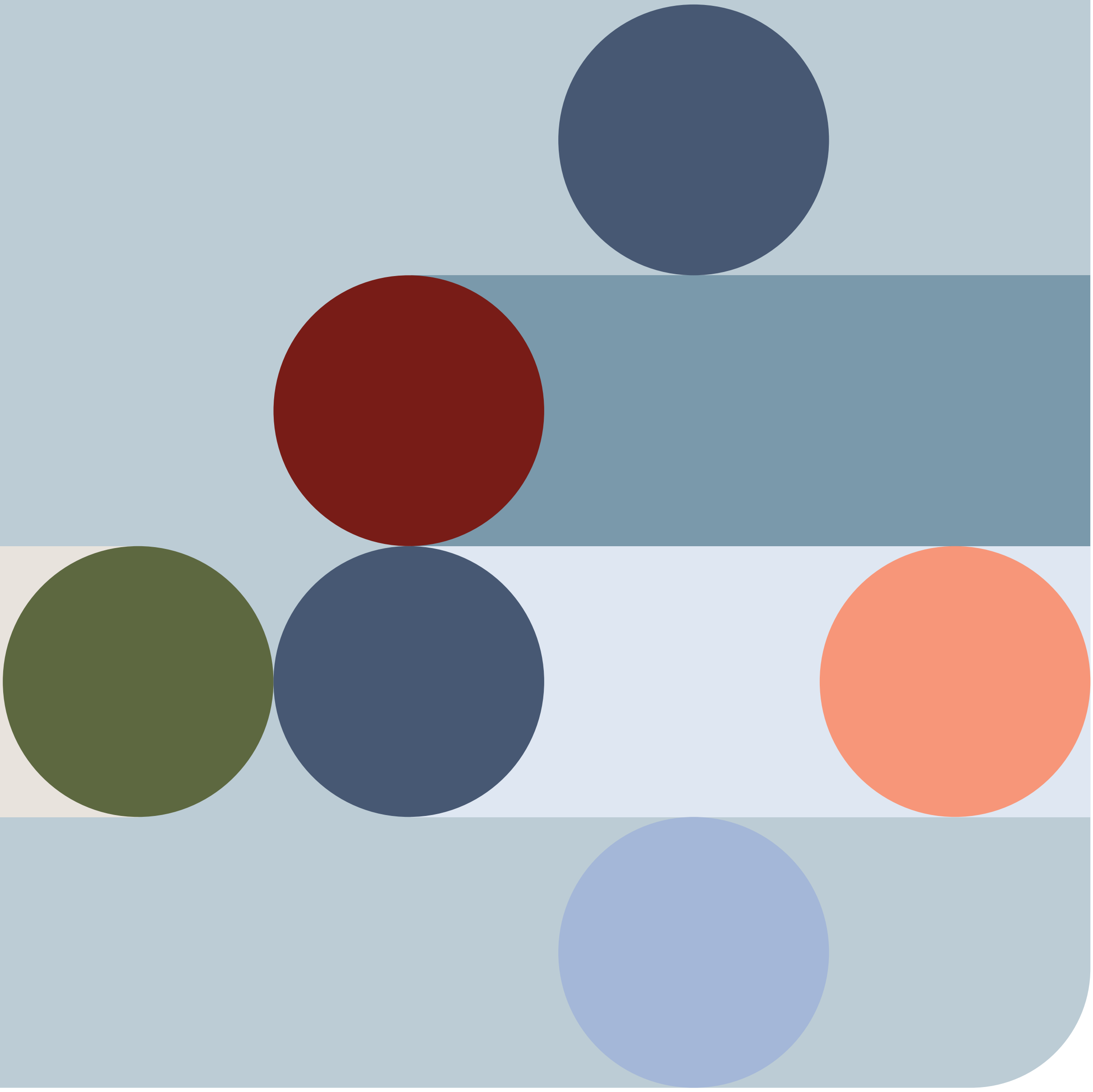
# Australian users' experiences with control features on social media services and online dating apps

Key findings

May 2023



MONASH
University

**Acknowledgement of Country**

The authors acknowledge the Wurundjeri people of the Kulin Nations on whose unceded lands Monash University operates. We pay our respects to Aboriginal and Torres Strait Islander Elders past, present and emerging. We value Aboriginal and Torres Strait Islander histories, cultures and knowledge.

**Funding acknowledgement**

This project and Final Report were commissioned by the Australian Government Department of Infrastructure, Transport, Regional Development, Communications and the Arts.

**Suggested citation**

Flynn, A., Wheildon, L., Robards, B., Vakhitova, Z. & Harris, B. (2023) *Australian users' experiences with control features on social media services and online dating apps*: *Key findings*. Report Prepared for the Department of Infrastructure, Transport, Regional Development, Communications and the Arts.

# Contents

## List of figures and tables

# Executive summary

## Project background and methodology

This Report details the findings of a study that examined Australians' knowledge, use and perceptions of the effectiveness of user control features (e.g., safety tools) on social media services and online dating apps. It also explored Australians' awareness of third-party solutions for preventing, reporting, and responding to online harm.

The study used a child and trauma-informed approach. It consisted of a walkthrough analysis of 20 social media services and online dating apps, 24 online focus groups of one-hour duration with 102 Australians aged 13-74 years, and two one-on-one online interviews with Australians aged 18+ years of 45 minutes duration.

### Research questions (RQs)

**RQ1** What social media services and online dating apps are most used by Australians?

**RQ2** What user control features are available (or advertised as available) to Australian users of social media services and online dating apps, and what knowledge do Australians have of these user control features, including third-party solutions?

**RQ3** Do Australians use control features, including third-party solutions to control their experiences when using social media services and online dating apps?

**RQ4** How do Australian users perceive the effectiveness of control features that they know of and use?

# Key findings

## Commonly used social media services and online dating apps

- Amongst research participants, the top six social media services used were: Facebook (74%), YouTube (69%), Instagram (69%), LinkedIn (47%), TikTok (37.5%) and Twitter (36.5%). The top six online dating apps used were: Tinder (27%), Bumble (21%), Grindr (8%), Hinge (6%), eHarmony (6%) and Plenty of Fish (2%).
- These rates are relatively representative of social media service and online dating app usage patterns in other quantitative studies of the general population in Australia.
- Australians' use of social media and online dating apps is increasing, but their feelings of safety are not.

## Available user control features

- The control features available on social media services and online dating apps varied significantly, but all had basic functions, such as the ability to block and report other users.
- Some social media services and online dating apps offered extremely nuanced levels of control (e.g., Facebook's custom sharing settings and manual user-driven algorithmic training). In contrast, others had more rudimentary, but still effective functions (e.g., Grindr's filter/favourite/block/report functionality).
- Some social media services and online dating apps have developed innovative and advanced control features for users, such as Tinder's 'Garbo' and 'Noonlight', however at the time of writing, these are only available for paying users located in the United States and are not accessible to Australians.

> **Some control features were positively perceived by participants because they helped users feel safer and more comfortable using the platform. It also suggested to them that the platform takes their safety seriously.**

## Awareness

- Participants had limited knowledge of the different control features and settings available on social media services and online digital dating apps, outside blocking and reporting functions.
- There was a lack of awareness among participants of third-party solutions.
- Participants expressed awareness of a range of online harms and risks associated with using social media services and online digital dating apps.
- The risk of experiencing harm is more significant for vulnerable and minoritised groups, such as Indigenous and First Nations people and LGBTQI+ people.

## Perceived effectiveness of user safety controls

- Some control features were positively perceived by participants because they helped users feel safer and more comfortable using the platform. It also suggested to them that the platform takes their safety seriously.
- The key effective control features identified by participants included: the ability to report fake profiles; policies that prevent racist or offensive language in online dating profiles; automated control features that prevent future profiles with the same contact information from following or friending users; the ability to block individuals across multiple accounts; AI functionality that blurs potentially offensive images; the ability to block contacts; filters that enable users to block words or hashtags they do not want to be exposed to in their feeds; and community notes, a crowd-sourced fact-checking program.
- Users are more likely to continue using a platform and recommend it to others when they consider its control features to be effective.

> **Australians' use of social media and online dating apps is increasing, but their feelings of safety are not.**

## Perceived ineffectiveness of user safety controls

- Reporting functions were perceived to be ineffective across the broad range of social media services and online dating apps used by participants for a range of reasons including: their complexity, the lack of information on how to access and use them, the limited nature of reporting categories (they don't always reflect the user's experience), and due to limited transparency around reporting outcomes.
- Digital platform responses to reports were perceived as ineffective and reduced participants' likelihood of reporting, due to responses often being delayed, disappointing or non-existent.
- **Some participants felt the reporting functions were not designed for them or their context, and there was a general sense of helplessness that users have to endure offensive content because reporting functions are ineffective. This was particularly relevant for participants reflecting on offensive content directed at vulnerable and minoritised communities, including racist, ableist and anti-LGBTQI+ community posts.**
- Participants, particularly those from the younger age focus groups, felt people could simply get around control features by creating new accounts or gaming the automated filters.
- Social media services and online dating apps' policies on how they define and respond to offensive or abusive content were viewed as inconsistent by participants, failing to protect vulnerable and minoritised communities and reflect social norms.
- Participants felt social media services and online dating apps depend too heavily on automated options (e.g., AI moderating tools) to assess reports of harmful content or behaviour, instead of involving human moderators, and thus overlook the context of offensive content.
- Social media services are not perceived to be enhancing the safety of children online.

## Suggestions

- **Social media services and online dating apps should ensure their policies are consistent, transparent and regularly reviewed and updated to reflect social norms and values, including those of vulnerable and minoritised groups. Updates must also be communicated to make users aware of new features or changes.**
- Increasing the involvement of human moderators, in addition to AI tools, to detect and respond to online harms on social media services and online dating apps would provide a more individualised and human-centred approach in assessing and addressing harmful information and behaviour.
- Social media services and online dating apps should provide updates on the outcome of any reports made and provide links to local support services and information beyond the platform, making sure these match the geographical location of the user.
- Social media services and online dating apps could consider expanding the scope of the control features that already exist on their platforms elsewhere, for example, in the United States, to Australian users.
- Control features should be customisable, so users have autonomy over the features to allow the user to filter out any content they personally consider to be offensive. However, there should be basic default standards for social media services and online dating apps to filter out blatantly harmful content automatically.
- Social media services and online dating apps should have an 'opt-out' mechanism for certain default control features, rather than users having to locate and 'opt-in'. However, these mechanisms need to be adjusted to the context and social conventions of the platform, rather than one-size-fits-all. Each platform can have a different user-base, so control features need to be developed in the context of differing user conventions.
- Social media services should do more to confirm users' ages and ensure age-appropriate content.
- Social media services should have more onus on them to engage with young people and make them aware of safety measures, including control features.

- Social media services and online dating apps could do more to improve processes to confirm users' identities where it is appropriate to do so. We acknowledge that anonymity can be useful and productive on some platforms.
- Social media services and online dating apps should utilise opportunities to raise awareness of the different functions, settings, and guidelines of control features through in-app prompts and reminders, drawing, for example, on common user-nudging practices already used on platforms.

## Conclusions

**More education on control features, using real case studies, is needed among young people in schools and through more widespread messaging to the general public to ensure Australians better understand the available online safety and control features, and resources. Organisations like the eSafety Commissioner can play a significant role in achieving this by sharing online safety information and resources.**

The lack of knowledge and confidence around third-party solutions and laws among Australians further emphasises the need for accessible and easy-to-understand information being more widely available to digital platform users.

Social media services and online dating apps should prioritise online safety and invest in developing and implementing efficient user control functions and automated control features, as well as improved education and awareness campaigns tailored for Australians from diverse communities, including Indigenous and First Nations people, culturally and linguistically diverse people, LGBTQI+ people and people with a disability.

The broader Australian response to online safety should expect social media services and online dating apps to improve user awareness of control features.

# Project and Report Overview

Helping Australians stay safe online and reducing online violence against vulnerable community members is a priority for the Australian Government. This can be seen in the expanded role of the eSafety Commissioner (*Online Safety Act 2021* (Cth)), who works with all sectors to assist and support those at risk of or experiencing online harms and the *Basic Online Safety Expectations* (BOSE), which sets out the Government's expectations of online service providers to take steps to keep Australians safe online. The online safety of Australians also forms part of the *National Plan to End Violence Against Women and Children 2022 – 2032*.

This Report details the findings of a study that addresses a pressing research gap on Australians' understanding, awareness and use of control features on social media services and online dating apps. The study examined Australians' knowledge and use of social media services and online dating apps, particularly their engagement with user control features (e.g., safety tools) for preventing, reporting, and responding to online harm. This investigation included examining Australians' perceptions of the effectiveness of control features available on social media services and online dating apps, and third-party solutions they had knowledge of, designed to enhance online user safety. A better understanding of how these features and tools are understood and used is vital for all Australians, but especially for vulnerable and minoritised populations, including women, Indigenous and First Nations people, LGBTQI+ people and culturally and linguistically diverse communities. This focus is necessitated because these groups experience harm, harassment, and violence at greater rates than the general population, and their voices and experiences are vital in understanding the role of control features and how they can be improved.

The study discussed in this Report was explicitly designed to generate direct and workable evidence for government policymakers, social media services and online dating apps and members of the public. In this regard, it makes a timely contribution to an under-researched area of growing national (and international) significance.

## Research Questions

The project responded to four research questions:

RQ1 What social media services and online dating apps are most used by Australians?

RQ2 What user control features are available (or advertised as available) to Australian users of social media services and online dating apps, and what knowledge do Australians have of these user control features, including third-party solutions?

RQ3 Do Australians use control features, including third-party solutions to control their experiences when using social media services and online dating apps?

RQ4 How do Australian users perceive the effectiveness of control features that they know of and use?

## Report structure

This Report presents the key findings from the focus groups, interviews and walkthrough app analysis. It begins with an Executive summary and outline of the project aims and methodology. It then provides an overview of current research relevant to online safety and user control features and presents data on which social media services and online dating apps are most used by Australians (RQ1). The Report then moves onto discussing what control features are available to Australian users of social media services and online dating apps (RQ2) and their knowledge and use of control features (RQ2 and RQ3), before presenting data on how Australians perceive the effectiveness of user control features (RQ4) and whether their experiences with control features align with what social media services and online dating apps purport to do. The Report concludes with a summation of the key findings and implications from the study.

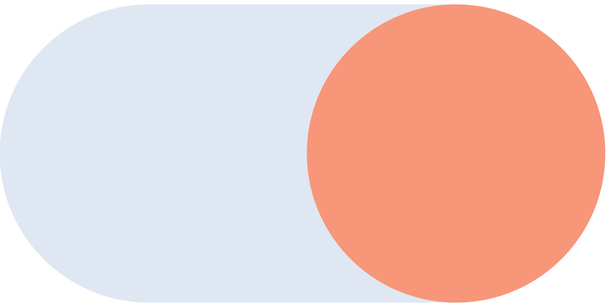
## Methodology

The study consisted of a walkthrough analysis of 20 social media services and online dating apps, 24 online focus groups with 102 Australians aged 13-74 years and two one-on-one online interviews with Australians aged 18+ years.

## App walkthrough

The 'app walkthrough method' developed by Light, Burgess and Duguay (2018; see also Møller & Robards, 2019) involves systematically and forensically stepping through the various stages of app registration and entry, as well as everyday use and disconnection practices, with a focus on control functions such as blocking, muting, unsubscribing, unfollowing, curating timelines, filtering hashtags and terms. While this method can be undertaken through web browser interfaces, we focussed on smartphone app interfaces.

The process involved listing and documenting function and control tools to generate knowledge of the control features on 20 of the most popular social media services and online dating apps' amongst Australian users. Building on the parameters set out by Light et al. (2018), we developed a 13-step protocol that mapped out: account creation, profile creation, user interface, main feed, feed curation, profiles, managing other users, rules, support and help, tone and experience of support, other platform-level control settings, profile/account suspension, and profile/account deletion. Our focus was on mapping user control functions to answer RQ2, but the walkthrough method allowed us to understand platform use more holistically in order to locate control functions within a wider constellation of use and platform context.



## Focus groups and one-on-one interviews

Twenty-four focus groups of one-hour duration and two one-on-one interviews of 45 minutes duration were conducted across Australia via Zoom over January and February 2023, with a total of 104 participants aged 13-74 years. The one-on-one interviews were conducted with participants who self-identified as having a disability or condition that made it easier to participate in an individual, rather than a group setting. Ethical approval to conduct the research was received from the Monash University Human Research Ethics Committee (Project No: 36480).

Participants aged 16+ were recruited through paid social media advertisements on Twitter, Instagram and Facebook. Over 800 people expressed an interest in participating. Participants aged 13-15 years were recruited with the assistance of a project recruitment company. To capture a sample broadly reflective of the Australian population, all people (regardless of age) who expressed an interest in the project were asked to complete a demographic questionnaire. The questionnaire asked potential participants about their age, gender identity, cultural background, sexuality, Indigeneity, socio-economic status and state of residence. The selection process was informed by Australian census data in order to approximate a cohort of participants broadly representative of the Australian population. A total of 120 people were invited to partake in the project; 104 participated. Participants were provided a small honorarium ($40) to acknowledge their time. Table 1 presents the socio-demographic characteristics of the participants.

## Table 1. Socio-demographic characteristics of participants

| Characteristic | N | % |
|---|---|---|
| **Age (years)** | | |
| 13 – 15 | 20 | 19 |
| 16 – 24 | 16 | 15 |
| 25 – 39 | 40 | 38 |
| 40 – 54 | 16 | 15 |
| 55 – 74 | 12 | 11 |
| **Gender** | | |
| Female | 70 | 67 |
| Male | 31 | 30 |
| Non-binary | 3 | 3 |
| **Sexual orientation** | | |
| Straight (Heterosexual) | 56 | 54 |
| Bisexual | 16 | 15 |
| Queer | 3 | 3 |
| Gay or lesbian (Homosexual) | 6 | 6 |
| Unsure | 2 | 2 |
| Unspecified | 21 | 20 |
| **State/Territory** | | |
| New South Wales | 38 | 37 |
| Victoria | 38 | 37 |
| Queensland | 10 | 9 |
| Western Australia | 7 | 7 |
| Australian Capital Territory | 5 | 5 |
| Southern Australia | 4 | 4 |
| Tasmania | 1 | 1 |
| Northern Territory | 1 | 1 |

| Characteristic | N | % |
|---|---|---|
| **Indigenous and Torres Strait Islander Status** | | |
| Indigenous Australian | 10 | 9 |
| Torres Strait Islander | 1 | 1 |
| **Country of birth** | | |
| Australia | 60 | 58 |
| India | 9 | 9 |
| Malaysia | 3 | 3 |
| New Zealand | 2 | 2 |
| Other | 9 | 9 |
| **Index of Socio-Economic Advantage and Disadvantage**[1] | | |
| Quintile 1 (most disadvantaged) | 5 | 5 |
| Quintile 2 | 11 | 10 |
| Quintile 3 | 22 | 21 |
| Quintile 4 | 24 | 23 |
| Quintile 5 (most advantaged) | 39 | 38 |
| Unspecified | 3 | 3 |
| **Cultural background** | | |
| Oceanian[2] | 34 | 33 |
| North-West European[3] | 18 | 17 |
| South-East Asian[4] | 11 | 11 |
| Southern and Eastern European | 11 | 11 |
| Southern and Central Asian | 7 | 7 |
| North African and Middle Eastern[5] | 3 | 3 |
| North-East Asian[6] | 1 | 1 |
| Mixed | 14 | 13 |
| Unspecified | 5 | 5 |

[1] This information is based on the post code data provided by our participants and derived from Census of Population and Housing: Index of Relative Socio-economic Advantage and Disadvantage (IRSAD) Socio-Economic Indexes for Areas (SEIFA) Australia 2016 (Australian Bureau of Statistics [ABS], 2016).

[2] Includes Australian and New Zealand peoples.

[3] Includes British and Irish

[4] Includes Mainland and Maritime South-East Asian peoples.

[5] Includes Arab and Jewish peoples and people of the Sudan.

[6] Includes Chinese Asian peoples.

Of the participants, 67% identified as female, 30% as male and 3% as non-binary. Over half (54%) identified as straight (heterosexual), with 15% identifying as bisexual, 6% as gay or lesbian (homosexual), 3% as Queer, 2% as unsure about their sexual orientation and 20% preferring not to say. The participants were primarily born in Australia (58%), with the next most significant majority being born in India (6%), Malaysia (3%) and New Zealand (3%). Participants were also born in Iran, England, France, Greece, Hong Kong, Lebanon, South Africa, Spain and Sri Lanka. Approximately 10.5% of participants identified as Indigenous or First Nations (Aboriginal and Torres Strait Islander). Participants came from a wide range of cultural backgrounds and socio-economic situations, as outlined in Table 1 above. In terms of age, 19% were aged between 13-15 years, 15% between 16-24 years, 38% between 25-39 years, 15% between 40-54 years, and 11% between 55-74 years. At the time of completing the focus groups, most participants lived in NSW (37%) and Victoria (37%), followed by Queensland (10%), WA (7%), the ACT (5%), SA (4%), Tasmania (1%) and the NT (1%).

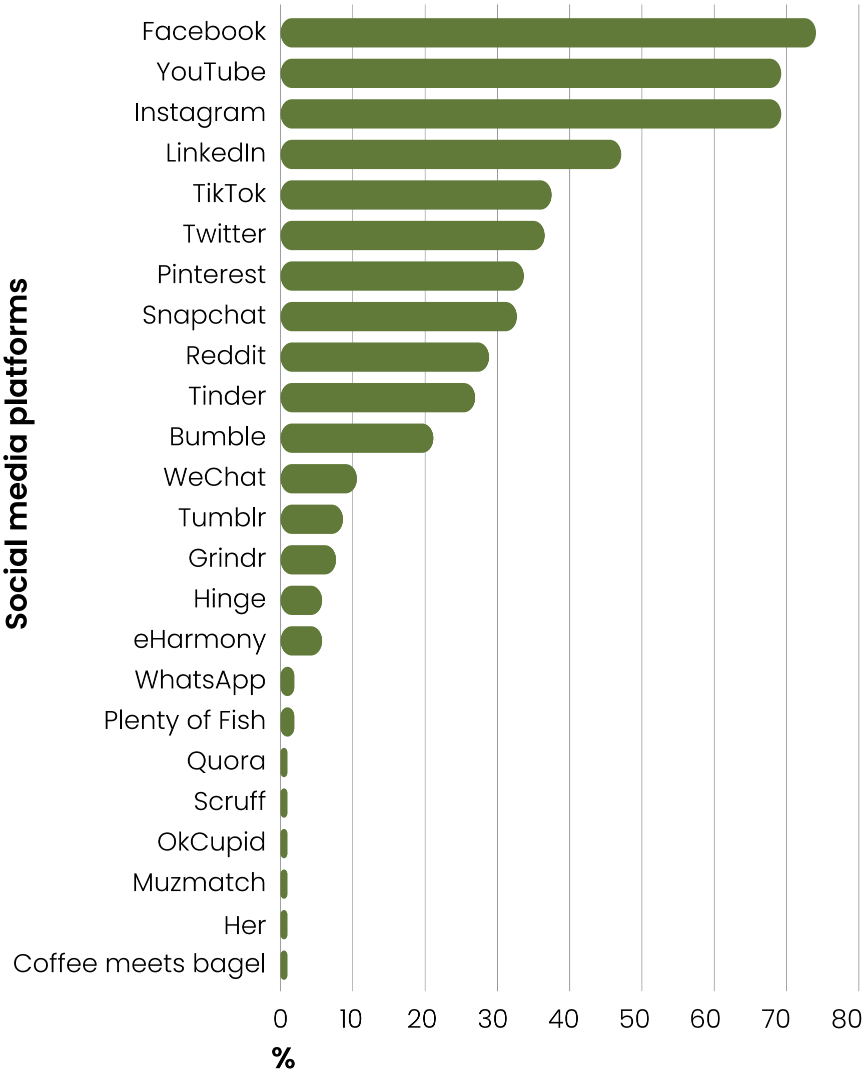
The focus groups were conducted in accordance with best practice guidelines in the field (WHO 2016), embodying a sensitive and considerate framework that prioritises participants' well-being and safety. Two members of the team conducted each focus group. One member of the research team conducted the one-on-one interviews. All focus groups and interviews were audio recorded and transcribed verbatim. Some of the quotes presented in this report have been slightly edited to remove hesitations (e.g., um, ah, like) or repetition (e.g., a, a, a digital platform), but not to distort their meaning. Identifying information (real names, specific locations, businesses, etc.) has been removed to maintain participant anonymity. Throughout this report, only the focus group or interview identifier is provided. This includes the focus group (FG) or interview number, e.g., FG16, FG2 or Interview 1, the composition of the group, e.g., 13-15 years, LGBTQI+ or 25-39 years, and the gender composition of the group, e.g., female, male, non-binary, or mixed (i.e., female, male and non-binary). For the two interviews, even though these involved one participant per interview, an age range is provided to reduce the potential for identification.

The focus group data was analysed using Dovetail – a qualitative data analysis platform that allows research teams to code and analyse data together in real time across devices. All team members were engaged in the thematic analysis process, which involved developing a set of codes relevant to the four RQs and analysing data according to these codes. Key trends were then identified and are presented in this Report.

# Australians' use of social media services and online dating apps

In 2022, smartphone ownership was at 92.2% among Australian internet users aged 16-64 years, and between 2022 and 2023, the number of mobile connections increased by 1.2 million (3.9%) to 32.71 million (Kemp, 2023).[7] Recent statistics indicate that the number of users of social media services continue to grow and that there are currently around 21.3 million active users in Australia, equating to 81% of the total population (Kemp, 2023). As outlined in Figure 1 below, there were 24 different social media services and online dating apps used by the focus group participants (RQ1).

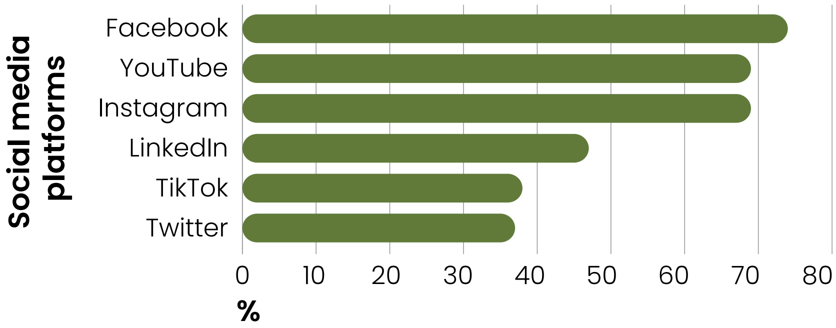
**Figure 1. Participants' use of social media services and online dating apps**



[7] The total number of mobile connections thus exceeds the total Australian population of 26.31 million.

Of these, the six most popular social media services used by the focus group participants, as outlined in Figure 2 below, include: Facebook (74%), YouTube (69%), Instagram (69%), LinkedIn (47%), TikTok (37.5%) and Twitter (36.5%).

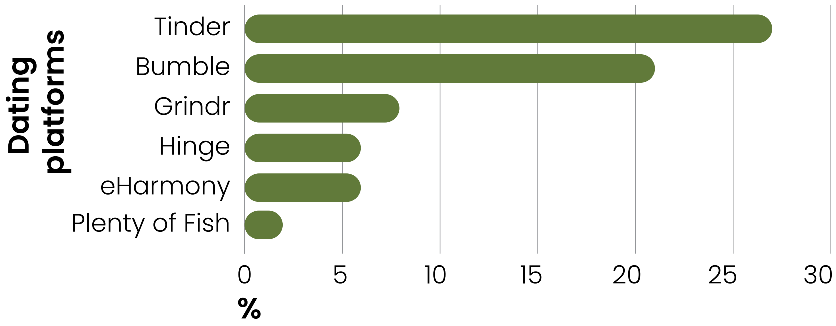
### Figure 2. Most popular social media services used by participants



The significant uptake in the use of smartphones globally has also seen a corresponding increase in the creation and number of online dating apps, which provide users with opportunities for friendship, connections, relationships and sexual intimacy (Duguay, 2020; Robards & Lincoln, 2020; Van De Wiele & Tong, 2014). Research has found that meeting partners online is becoming the most common way for couples to connect (Rosenfeld et al., 2019). According to a recent Statista report, over 3.2 million Australians (2.58 million non-paying and .68 million paying) were actively seeking partners using online dating apps in February 2023. This figure is forecast to increase to 3.4 million users by 2027 (Statista, 2023).

In January 2023, the top six most downloaded free online dating apps across all stores in Australia were: 1. Bumble 2. Hinge 3. Tinder 4. Wink 5. Plenty of Fish and 6. Hily. The five top-grossing apps were: 1. Tinder 2. Bumble 3. Hinge 3. Grindr 4. Zoosk and 5. Plenty of Fish (AppMagic.Rocks, 2023). These figures largely correspond with the most common online dating apps used by participants in our study. As outlined in Figure 3 below, the six most popular dating apps among the focus group participants included: Tinder (27%), Bumble (21%), Grindr (8%), Hinge (6%), eHarmony (6%) and Plenty of Fish (2%).

### Figure 3. Most popular online dating apps used by participants



## Online harms

The growth in online dating apps and social media services has seen a corresponding increase in concern about online safety among Australians, with a recent eSafety Commissioner (2022) survey of adults aged 18-65 years finding that:

- 75% of Australians had experienced something negative online, up from 58% in 2019;
- Just under 1 in 3 (33%) Australians said their negative online experiences impacted their emotional and mental well-being; and
- Almost 1 in 6 (17%) Australians said their negative experiences online affected their physical health.

In relation to online dating apps, Australian research indicates that in addition to having positive outcomes in connecting people, they can facilitate digital dating abuse – a form of interpersonal violence, harassment or abuse perpetrated in the context of a current, prospective or former dating relationship. Digital dating abuse can involve 'real world' contact, with violence potentially being perpetrated both 'online' and 'offline' and in various relational settings (Gillett, 2021; Flynn et al., 2022). A recent study in Australia (Wolbers et al, 2022) found 3 in 4 (75%) people have experienced sexual violence due to using online digital dating apps. Further, Rowse et al.'s (2020) review of forensic sexual assault investigations in Victoria found 14% of sexual assault victims (all of whom were women) were assaulted by a perpetrator after meeting on a dating app.

Fake profiles and the lack of age and identity confirmation required to create accounts on social media services and online dating apps have been identified in research as somewhat fraught, with underage access to material that is not age appropriate, and fears about child predators and grooming being balanced against privacy and data protection considerations regarding the collection of personal information, particularly of minors (Thierer, 2007). These issues are nonetheless growing concerns for Australian users with the eSafety Commissioner's (2022) study revealing a significant increase in the number of people who have had someone pretending to be them online (from 9% in 2019 to 16% in 2022).

## Vulnerable and minoritised communities

Research indicates that particular population groups, including women (Flynn et al., 2022; Harris & Woodlock, 2021), Indigenous and First Nations people (Carlson, 2020; Carlson & Frazer, 2021; Flynn et al., 2022), LGBTQI+ people (Byron et al., 2019; Nelson et al., 2022) and culturally and linguistically diverse communities (Henry et al., 2022) are disproportionately subjected to technology-facilitated abuse, discrimination and marginalisation. Research further shows that online abuse is gendered, with cisgender men overrepresented in data regarding perpetration and cisgender women and LGBTQI+ people overrepresented as victims (Brown et al, 2021; Cama, 2021; Carlson, 2020; Duncan & March, 2019; March et al., 2021; Rowse et al., 2020). It has been argued that in this context, the absence of adequate or trusted safety mechanisms means vulnerable and minoritised communities have to undertake extensive 'safety work', investing time, energy and resources into efforts and strategies to prevent violence and protect themselves (Gillett, 2021). This may include being pressured into or electing to disengage from technology (Harris & Woodlock, 2021). There is also an emerging body of research from First Nations researchers in Australia demonstrating that Indigenous social media service and online dating app users regularly witness and experience racism, sexism and 'sexual racism' online (Carlson, 2020; Kennedy, 2020). In their review of the literature on cyberbullying and Indigenous Australians, Carlson and Frazer (2018) found that social media helps proliferate hate speech, including racism and other forms of violence. Matamoros-Fernández (2017) similarly found that the prioritisation of free speech on social media services such as Facebook, resulted in them favouring "offenders over Indigenous people" (p. 931).

Carlson (2021) contends that without a strong incentive for social media services to tackle the issues of hate speech and racism, they will not do so. However, Carlson and Frazer (2021: pp. 250-251) also reveal that "Indigenous peoples are engaged complexly, adeptly and playfully with digital technology" and that "the connections sustained online are not just sites of social tension, but are very often productive of joyful, experimental social formations of care, desire, friendship and love".

Much literature examining safety on social media services and online dating apps highlights that technologies are not created in a vacuum. Technology bodies and social media services and online dating apps tend to be monocultural, which means that cultural assumptions, values and ideas are often unintentionally built into hardware, software, digital cultures and control features (Chang, 2018; Ionescu, 2012; Reed, 2018; Suzor, 2019). Existing control features have also been critiqued on the grounds that they can have unintended consequences and enable increased surveillance and criminalisation, especially of vulnerable groups (Stardust et al., 2022). Concerns around the monocultural nature of social media services and online dating apps have led to some major changes in the composition of the workplace. Meta, for example, has identified increasing representation in their workforce as a priority, setting goals to: double the number of women employees globally (to at least 50% of their workforce) and the number of Black and Hispanic employees in the US; increase the number of leaders (director-level employees and above) who are people of colour by 30% in the US; increase the number of people from underrepresented minorities, including people with two or more ethnicities, people with disabilities and veterans in the US (Meta, 2021). It is hoped by making changes like this, social media services and online dating apps can be more responsive to the diversity of their users, and better address key areas of concern specific to minoritised and vulnerable communities' experiences of harm online.

In sum, Australians' use of social media and online dating apps is increasing, but their feelings of safety are not and the risk of online harm is more prominent for vulnerable and minoritised groups. The eSafety Commissioner's (2022) study found that Australians want technology companies to do more to keep them safe. The study found that:

- 82% say tech companies have a responsibility for their online safety;
- 42% say tech companies aren't doing enough to build control features into their services and products;
- 58% want safety and privacy settings set to the highest setting by default;
- 57% want user content to be scanned to detect illegal or seriously harmful content so it can be removed; and
- 51% want tools to report inappropriate content.

While this shines light on what Australian users want from social media services and online dating apps, there is limited research on what Australians, and vulnerable and minoritised groups in particular, know about user safety control features, how they use them and whether they consider these to be effective in protecting them online. In the next section, we report findings on what knowledge Australian digital platform users in our study have of user control features and whether and how they use these features.

# Available user control features

To better understand what user control features are available to Australian users of social media services and online dating apps (RQ2), we undertook an 'app walkthrough' (Light et al., 2018) of the control features on 20 of the most popular social media services and online dating apps' amongst Australian users. We focused on the smartphone app interfaces and control features, noting that some functions appear differently on web interfaces. Our comprehensive analysis identified a wide range of control features available on these services and apps across three key areas: (1) platform-level features, (2) user-level features, and (3) third-party solutions.

## 1. Platform-level features

a. Banning users (for violating platform terms of service, often through user reports covered below in 2C).

b. 'Shadowbanning' or visibility throttling (such as algorithmic de-prioritisation of #BlackLivesMatter content on TikTok).

c. 'Quarantining' or deleting groups (such as Reddit quarantining offensive subreddits or Facebook banning hate groups).

d. Account verification (usually an optional feature, such as 'Meta Verified' where government issued ID is used to establish an account's 'authenticity').

e. Pre-emptive image blurring on potentially sensitive or explicit content (for example, Bumble's 'Private Detector' tool).

## 2. User-level features

a. Blocking other users.

b. Pre-emptive blocking (using tools such as Tinder's 'block contacts' function that allows users to pre-emptively block contacts they do not wish to see, like siblings, friends, co-workers, etc.).

c. Reporting users or individual posts.

d. Unfriending/unfollowing contacts.

e. Muting users or specific content (such as muting a contact on Instagram, so their stories do not appear in the story feed or muting certain words or hashtags on Twitter like 'wordle' or '#TheBachelorAU').

f. Filtering users (on dating apps, for instance, by demographic characteristics such as age, height, or 'tribes' on Grindr) or content (such as TikTok's 'restricted mode', Facebook's 'see less content like this' on certain posts, or Instagram's 'sensitive content control' tool).

g. Using differentiated privacy settings for sharing (e.g., posting to specific groups on Snapchat, 'circles' on Twitter, 'close friends' on Instagram, or custom sharing/sharing with certain lists of contacts on Facebook).

h. Add context functions (such as the 'community notes' function on Twitter, allowing approved contributors to add context to potentially harmful tweets).

i. Maintaining multiple profiles (such as 'throwaway' accounts on Reddit or swapping between accounts on Instagram).

j. Setting profiles to 'public' or 'private' (a simple binary on Twitter and Instagram) or more customisable sharing settings (such as on Facebook with different parts of profiles being shared or kept private selectively, and at multiple levels).

k. Temporarily deactivating an account or profile (such as 'hibernating' a LinkedIn account or deactivating a Facebook account).

l. Permanently deleting an account.

## 3. Third-party solutions

a. The eSafety Commissioner, Australian Centre to Counter Child Exploitation (ACCCE) or police reports. The eSafety Commissioner investigates cyberbullying of children, adult cyber abuse, image-based abuse (sharing, or threatening to share, intimate images without the consent of the person shown) and illegal and restricted content. The eSafety Commissioner is empowered to administer different reporting and takedown schemes in relation to image-based abuse and 'seriously harmful content', under the *Online Safety Act 2021* (Cth). The ACCCE investigates reports of online sexual extortion or blackmail, the use of technology to facilitate the sexual abuse of a child, production or sharing of child sexual abuse material online. If someone is in immediate danger, Triple Zero (000) and the police should be contacted.

b. Parental control or screen time monitoring via apps, or Wi-Fi and phone operating system functions.

c. Additional app-linked 'safety partnerships' (for example, Tinder's 'Garbo' and 'Noonlight' functions).

d. Apps such as followerAudit and inbeat, which help users identify fake Twitter and Instagram followers, and help marketers verify influencers' audience.

Our walkthrough analysis of social media services and online dating apps shows that there are a broad range of control features available, however, some of the more innovative control features are only available to users in the United States. For example, on Tinder, there is a paid feature called 'Garbo', which provides users with access to a 'background check platform that allows you to search public records including arrests, convictions, and sex offender registry information to help you feel safe' (Tinder Safety Centre, 2023). Similarly, the feature entitled 'Noonlight' provides users with a 'backup every time you meet with someone'. This means that users can share 'where, when, and who you're meeting IRL [in real life]', creating a record of plans to meet other users in-app from Tinder to Noonlight (Tinder Noonlight FAQs, 2023). You can also 'signal for emergency help' by pressing a button that will notify local police. Similarly on some social media services and online dating apps, the main resources provided were almost entirely based in the US, for example linking to US-based sexual assault hotlines (RAINN) and the US National Domestic Violence Hotline, and it was more challenging to locate localised Australian resources, such as 1800 RESPECT, Lifeline, QLife and Women's Services Network. **This suggests there is scope for social media services and online dating apps to enhance the control features available to Australian users, and for them to ensure localised resources are easily accessible based on the geographical location of the user.**

# Australians' knowledge of user control features

Having established the types of control features available on social media services and online dating apps used by participants, we next sought to examine their knowledge of the various features, and then ascertain their views on the effectiveness of these features. We begin by exploring their knowledge of control features and then exploring their use of these features.

## The most commonly known control features

Focus group participants had a range of levels of awareness and insights regarding user control features relating to online harm and safety on social media services and online dating apps (RQ2). The three most commonly known control features identified by participants were reporting and blocking, followed by the privacy settings that keep user profiles accessible only to those individuals/groups the user permits to follow/friend them and which can prevent strangers from sending direct messages. The below sample of comments is representative of the common responses participants provided when asked what control features they had knowledge of:

***The report and the block buttons***
(FG16: LGBTQI+, 27-30 years, male).

***You can report or block other people***
(FG23: 13-15 years, mixed).

***Block them and report***
(FG24: 13-15 years, mixed).

***Report and block***
(FG9: 32-39 years, female).

***There's a process of reporting and blocking***
(FG4: LGBTQI+, 27-44 years, mixed).

For some participants, reporting and blocking seemed to have become almost a mantra:

***Yeah, definitely a block, a report, an un-match, all of the above***
(FG5: 26-52 years, female).

Many research participants also referred to knowing about control features or settings that keep user profiles private and prevent connection requests or messages from strangers:

***I know about the Telegram feature, whereby you can have some restrictions to yourself, avoiding requests from strangers and also avoiding people putting you into a public group that you do not actually give consent to***
(FG15: LGBTQI+, 22-39 years, female).

Some participants knew of tools that allow users to 'train algorithms' through options such as 'hide post', 'see fewer posts like this', 'not interested in this Tweet', and options to 'snooze' or 'mute' an account either temporarily (e.g., 'temporarily stop seeing posts for 30 days') or permanently. As these participants observed:

***Lots of the platforms have a way to sort of say that you are not interested in this or don't wanna see it and sort of trying to take the effort to kind of train it away from that***
(FG13: LGBTQI+, 18-47 years, mixed).

***I do use the feature pretty regularly, the "I don't want to see content like this" on Instagram, particularly around weight loss and exercise, and food, stuff like that***
(FG5: 26-52 years, female).

***My "For You" page, the algorithm is pretty perfect for me, based on the images that I want to see***
(F2: LGBTQI+, women and non-binary).

## Five common user control features

Further to reporting, blocking and privacy settings, there were five other main types of control features raised as examples across more than one focus group. These included:

**1. Instagram allowing users to block individuals across multiple Instagram accounts** (i.e., to block an individual's account and other existing accounts they may have or accounts they may create):

*I know Instagram has, if you block someone on there, it comes up with, "do you want to block them and sort of any future accounts they make?". I think it's tied to the device or some sort of identifier that they have. I haven't seen that elsewhere*

(FG13: LGBTQI+, 18-47 years, mixed).

*I know that's an option in Instagram, you can block someone and other accounts that might be created from, I guess it's the same email address*

(FG15: LGBTQI+, 22-39 years, female).

**2. AI functionality on social media services and online dating apps which blurs potentially offensive images and provides users with the option of seeing them or not:**

*When a picture is a kind of sensitive one, it [Bumble] let[s] me know this photo is a kind of sensitive [one], so it gives me an idea of what I'm about to see. So, I think those features are very important*

(F3: LGBTQI+, mixed).

*I'm not sure if it was on Facebook or something, there was a warning before, it says something like, "This video contains graphic content. Do you wish to proceed?" And I've seen something like that … it's a good idea.*

(FG2: LGBTQI, 17-24 years, female and non-binary).

**3. The ability to pre-emptively block contacts on Tinder so you don't accidentally encounter them on the dating app:**

*When I was using it [Tinder]… there was an option like that [to block contacts]. I can't remember what the exact wordings are, but you could actually filter out your own friends to make it less awkward*

(FG5: 26-52 years, female).

This function was also identified as particularly useful for those escaping abusive relationships:

*It'd be useful, especially if you're in a DV [domestic violence] type situation and you are wanting to get away from people*

(FG16: LGBTQI+, 27-30 years, male).

**4. AI tools that allow users to train algorithms to see less of certain types of content on social media services and online dating apps. For example, filters that enable users to prevent words or hashtags they don't want to be exposed to from appearing in their feed:**

*I'm really familiar with TikTok. So, something that this person could do would be to blacklist the term, specific terms that they don't want to see about, like in that instance, it would be something like, homophobic or something like that. They could also report content like this, and then it could get rid of content like that*

(FG2: LGBTQI+ 17-24 years, female and non-binary).

*I don't know if it's on Instagram, but I know on TikTok there's this thing where you can filter out certain words in your comments*

(FG24: 13-15 years, mixed).

**5. Community notes on Twitter, which is a crowd-sourced fact-checking program that allows contributors to add context to potentially harmful Tweets:**

***Well in Twitter, we have a "community notes" and that's sort of a way of equalizing that sort of 25% of tightly wound people who have a very loud voice. It is sort of a way of equalising things that may do harm to people. You know, having a community of people who can add a note to something and not completely delete it from existence. I mean that still is pursuant to a real truth and that's important***

(FG13: LGBTQI+, 18-47 years, mixed).

## Accidental, learned or limited knowledge

Many participants reported that they became aware of control features by accident. For example, one participant described finding features by ***"accidentally pressing a button"*** (FG22: 13-15 years, mixed). Another described looking for something else and ***"finding it [the control feature] by accident"*** (FG7: 25-54 years, mixed). Other participants described finding out about features from friends and by word-of-mouth:

***It's only by word-of-mouth. I don't think I've seen it [how to report false profiles on dating apps] very clearly kind of advertised***

(FG18: 23-47 years, female).

***I had a friend that had a bad experience on Tinder because she met someone, and it wasn't safe. So, then she went through the process of reporting and then she told me that that was a thing coz I had never, I never really knew that you could do that. … So, then she kind of told her whole group that that's what's possible and to do that if something like that happens to us***

(FG18: 23-47 years, female).

***My friend messaged me and said, "This is actually a fake page [of their friend's profile], can you go and report it?". So, I think there's a lot of awareness, from my friend group, if this happens to me, what to do***

(FG3: 28-51 years, female).

Some younger research participants reported that some of their knowledge regarding online safety and safety resources came from the school curriculum and annual talks at school from police:

***In primary school, we got a bunch of assemblies with the police and stuff, they brought police in to talk about like, "don't share anyone your passwords", and stuff like that, but then going into high school, it was more what to do if you get bullied by people***

(FG21: 13-15 years, mixed).

***In school, because I'm like, I recently graduated … we had a lot of [visits] especially from police officers themselves who had come into school and tell [sic] us about cybersecurity***

(FG14: 18-40 years, female).

A few people mentioned an awareness of the eSafety Commissioner and how their resources informed their approaches online:

***I really love the resources which are on the eSafety Commissioner website. I use it a lot because I did attend one of their conferences, the first one, the inaugural one, and I really loved it, especially the animated characters which they have in the videos***

(FG3: 28-51 years, female).

***The eSafety Commission[er] was something we actually looked at specifically in our PDHPE [Personal Development, Health and Physical Education] class because … one module is on cybersecurity so looking at ways you can be more safe online***

(FG14: 18-40 years, female).

Others mentioned the importance of local groups, such as the Brisbane Lord Mayor's Youth Advisory Council, in sharing information about online safety.

An area where participants lacked knowledge and confidence in relation to online safety was third-party solutions. Many participants could not name any third-party solutions, beyond avenues such as police or other support/ reporting organisations. Others lacked trust or confidence in third-party options. As these participants observed:

***No, I don't know of anything like that***
(FG11: 35-40 years, mixed).

***I've personally never heard of anything like that***
(FG18: 23-47 years, female).

***You cannot always trust third-party solutions to necessarily fix your woes***
(FG8: 16-17 years, mixed).

Another area where participants expressed a lack of knowledge concerned laws:

***All those things are not very clear in terms of police and whatnot***
(FG16: LGBTQI+, 27-30 years, male).

***They need to create the laws around online safety and they need to make them to the extent that everybody understands what they are – not the full legalese jargon, but everyday speak, and people need to know what the laws are, because I can honestly say, right now, I wouldn't know what the laws are***
(FG6: 55-66 years, female).

# Australians' use of control features

## Lack of use

Focus group participants expressed mixed responses about whether they used control features on social media services and online dating apps (RQ3). Some reported that they do not use them:

***Honestly, I just like open it [TikTok], my kids got me onto it. I just open it and flick through the videos. I don't really go into the settings or anything like that***

(FG4: LGBTQI+, 27-44 years, mixed).

As discussed in more detail in the next section of this report, one of the primary reasons cited for not using control features was the perception that the features were ineffective:

***It depends on the platform what happens. Instagram, I find if I report something I might hear back a year later that we've looked into your post or to your report and this is the outcome. And so, it doesn't really make me feel very empowered as a user to report things***

(FG13: LGBTQI+, 18-47 years, mixed).

Another reason for not using existing control features was their complexity:

***I know there's a toolbar on Facebook, but it's quite extensive…. Yeah. But I wouldn't know how to use it all***

(Interview 2, 35-40 years, male).

***It's quite difficult to get things taken down***

(FG16: LGBTQI+, 27-30 years, male).

For some participants, there was a sense that the control features available were not created for them and, thus, did not help them. As discussed in more detail in the next section of this Report, this was raised by a number of participants from vulnerable and minoritised communities, including participants identifying as Indigenous and First Nations, participants with a disability and LGBTQI+ people. As this profoundly deaf research participant said:

***I left the Facebook singles dating group site for deaf Australians because it is full of toxic [sic] and impacts mental health wellbeing of deaf people. … I did report issues to social media sites, but they fell on "deaf ears"***

(Interview 1, 30-35 years, male).

One Aboriginal research participant said he had his own filter (let's call it the ***"Alex filter"***) and that he found the ***"Alex filter"*** more effective than anything else (Interview 2, 35-40 years, male). The ***"Alex filter"*** involved keeping accounts and posts private, identifying and avoiding bots, blocking messages from accounts he was not connected to and ignoring offensive comments:

***I can, you know, if I don't like a comment or whatever, I just, yeah. Whatever. That's that person's opinion. I don't need to agree with it***

(Interview 2, 35-40 years, male).

Other participants similarly advised that people must take matters into their own hands. In the context of receiving an unsolicited image, one focus group discussed the following option:

***P1: Maybe if they wanted to, they could probably post it in one of those Facebook or Instagram groups and I don't know if you've seen them, they just make fun of people really.***

***P2: Yeah, I've seen them before. So, like take a screenshot and then post it. … Like I've seen, "entertaining douchebags of Grindr"***

(FG16: LGBTQI+, 27-30 years, male).

Other participants felt less prepared to report and block people than others:

***I don't tend to block people too quickly unless it's sort of obvious that they will never have anything to contribute***

(FG13: LGBTQI+, 18-47 years, mixed).

***I do not know much about that, the reporting features. I think you can report issues, but I do see a lot of accounts that just look fake, I haven't reported them, but there are quite a few, I think. And then I haven't reported things on Twitter, harmful content. … And Facebook, I haven't had to, but I know a lot of people struggle with having fake accounts and fake accounts in their names***

(FG11: 35-40 years, mixed).

For participants who found some control features too complicated to navigate, blocking was considered a more straightforward control feature:

***It's quite difficult sometimes to find how to actually report a post and have that reported, as opposed to just being suggested to block someone***

(FG5: 26-52 years, female).

## Third-party solutions

A handful of participants had experience using third-party solutions. Reporting to an external authority, such as the Ombudsman, police or the eSafety Commissioner, was identified by several participants as a mechanism for responding to online abuse. As one participant explained:

***My [person they know] had a situation where someone was stalking her on Instagram to the point where she was being told where she was going, like someone was actually following her and taking photos of her and sending them to her. … I told her to take it to the eSafety Commissioner, and I think it went to the cops after that. … But that was our third-party solution. [But] … if something has happened in the sense of threats, abuse, sexting, all that sort of stuff, I've taken it to the eSafety Commissioner***

(FG5: 26-52 years, female).

Some participants identified using parental apps and third-party solution features to block non-age-appropriate content from their children's devices:

***I'm trying to think what I use just in my phone that I use for my kids. … I know that just with our router, like our internet service provider, I think they've also advised us. I think there's certain websites that we've blocked and with the kids' devices we can get alerts if they go to certain websites or block certain websites. … One of them is called Net Nanny, I think it's like a Google thing that I use that to try and control and just like supervise [the children]***

(FG4: LGBTQI+, 27-44 years, mixed).

Many women participants spoke about the 'safety work' they undertake when engaging online (Gillett, 2021). Recognising this, and the high levels of online harm women experience, some technology companies, social media services and online digital dating apps are implementing mechanisms and features to help address particular threats and contexts such as domestic and family violence perpetrators weaponising technology against victim-survivors. IBM, for example, has developed 'coercive control resistant design principles' (Nutall, 2020) and Apple has engaged domestic and family violence agencies in the modification and management of connected devices (WESNET, 2020). Further, representatives from Meta and Twitter have attended domestic and family violence sector events and engaged with practitioners in efforts to enhance the functionality and regulation of their social media services and online dating apps and increase women's safety (see WESNET, 2022). In 2023, WESNET (a non-government organisation that represents women's refuges, shelters, safe houses and referral services, nationally) collaborated with Tinder to create a Tinder Dating Safety Guide, to help survivors and broader community members safely use the app (WESNET & Tinder, 2023). Broadly, these initiatives and resources provide information about privacy and safety, which may include hiding or removing women's identifying information and digital trails, as a type of control feature.

Reflecting this 'safety work' some participants identified in-built control features on their digital technologies as a form of third-party protection. For example, one participant described the following:

***Apple has a few features like being able to hide your email when you're signing up for things, mainly like social media, being able to create strong passwords in terms of like hacking and things like that***

(FG14: 18-40 years, female).

Another participant also reflected on this feature:

***On iPhones now they've got a feature, it's in lab stage, where you can actually hide your email addresses – all communications through your phone, as well as have a virtual private network from the phone too – I think Apple organised it. You can actually opt into it if you go into your settings and then [it can] hide all your personal details and any identifying features when [you're] online***

(FG9: 32-39 years, female).

Others described using features that hide your location and encrypt messages, including apps and virtual private networks (VPNs) as a way to keep yourself safe:

***I use the app, which makes my location hidden … [and] apps that encrypts my information, my messages, and my location. I think that's my top priority, to hide my location. … So, I use this VPN, and I have to [pay] a very small monthly fee, but it does give me the security of hiding my location, and my messages, it encrypts my data***

(FG3: 28-51 years, female).

Related to third-party solutions for individual users, one participant described an example of a bystander safety campaign operating online that uses a hashtag to encourage others to post positive messages on content to compensate for hateful, abusive or offensive comments. This ***"cool Facebook group"*** (FG16: LGBTQI+, 27-30 years, male) #iamhere Australia was described as follows:

***It's pretty interesting, because what they do is they influence the tone that's on Facebook posts and on social media where there's something like there's a lot of racism, if you go on [their profile] … they write very positive messages and therefore influence what's happening. … You see people do the hashtag, like trans people on an ABC article or Channel Nine, they will, someone will post it here and say, "oh this post is not getting very nice friendly comments" and people will view that, young people, and they'll be affected by that. And you go and just write, "can you write something positive or do something?" [and they do]***

(FG16: LGBTQI+, 27-30 years, male).

# Effectiveness of user control features

## Effective user control features

Focus group participants drew on various personal examples to describe the effectiveness of user control features (RQ4). Many of the features identified as effective were on online dating apps. For example, one participant described their positive experience reporting a potential fake profile:

***I saw a guy's photo twice under two different profiles and I alerted them and … within half an hour I got a response. … And they did take it very seriously***
(FG5: 26-52 years, female).

Other users described control features and policies as being more effective when implemented by the platform, as opposed to being user-operated or requiring users to self-report an incident/issue. For example, one participant described a policy change on the dating app, Grindr, to prevent racist or offensive profiles from being created. They explained:

***They changed the policies on what you could put in your profiles, and they made it so that you couldn't have anything racist in there. I think for a long time, the typical thing … seeing a problematic Grindr profile, would be like "no fats", "no fems", "no Asians" … which is really awful, and you don't see that anymore because they made it really strict on what could be in the profiles***
(FG13: LGBTQI+, 18-47 years, mixed).

Some participants described feeling that control features developed and instigated by social media services and online dating apps themselves had a positive impact on the platform and were an effective way to make them feel safer using the platform. In the context of a dating app, one participant noted how a platform-initiated safety feature resulted in them feeling that the platform cared about user safety:

***On Bumble one day, I got a random message from a bot saying that they had removed a particular person from the site, and they were letting me know because I had friended them, or I had had contact with them. So that proactive response, I went yeah, okay, you are looking after me, you are telling me that there's been an issue, and you've done something. Now, no harm came to me, but it was reassuring.***
(FG5: 26-52 years, female).

Other examples identified by participants as being effective, outside of dating apps, included policies and automated control features that come into effect when you 'block' a person, which prevents any future profiles with the same contact information from being able to follow or friend you. As one participant observed:

***Currently, you can block someone and all future profiles that is connected to the same email, but before that, that was not a thing until maybe about two years ago. And I've had a few either perpetrators or bots trying to approach me online, and you try and block them, but they'll just make another profile, still the same email, and they'll just approach me again online. So, before that feature, I was having those sorts of problems***
(FG5: 26-52 years, female).

## Ineffective user control features

While participants identified some effective control features, most reported feeling dissatisfied with control features overall, reflecting that they did not meet their needs, or were ineffective in achieving the desired goal. Many of these concerns arose in the context of the reporting features available on social media services and online dating apps. As identified in our walkthrough analysis, all social media services and online dating apps examined in this research contain some reporting options. Indeed, reporting and blocking were the most consistently identified control features by the study participants, and participants had the most experience using these features.

The ineffectiveness of reporting features was not attributed to any particular platform, but rather something users felt was problematic in addressing online safety across social media services and online dating apps. Below, we outline some of the key comments around the ineffectiveness of control features with a specific focus on reporting functions.

## Lack of information

One of the key concerns identified by participants about the effectiveness of control features, and reporting in particular, was the lack of information on how to report inappropriate content or how to access other control features:

***It's quite difficult sometimes to find how to actually report a post and have that reported, as opposed to just being suggested to block someone or something like that***

(FG5: 26-52 years, female).

Another participant similarly observed:

***It just needs to be really easy to be able to report stuff. I don't know if you find out this information when you join a platform, … it should be more general knowledge. Like I'm learning so much just listening and now I want to go into the settings and stuff and change what I can see. But yeah, I don't know, like for people that aren't very computer illiterate, just reporting something on social media can seem really hard, really complicated***

(FG4: LGBTQI+, 27-44 years, mixed).

The absence of knowledge on what to do was similarly identified by another participant, who stated:

***You want to report it and you search [for how to report it], and you don't know what to do and you're flustered***

(FG14: 18-40 years, female).

Other participants reflected on the reporting process being ineffective in meeting users' needs. A key issue identified here was the options for reporting an incident, problem or situation, not adequately describing the user's experience. This concern was identified across social media services and online dating apps. As one participant observed:

***Well on Snapchat, you can just hold down the profile and report it, but sometimes the automated messages say why you want to report them, some of the reasons aren't on there***

(FG24: 13-15 years, mixed).

Other participants similarly described this experience, which then left them unable to report the behaviour:

***When you say you want to report something, the categories they've got, which I can't recall off the top of my head, but I know there's been a few times where I've wanted to report something, and there's just really not an appropriate option there***

(FG15: LGBTQI+, 22-39 years, female).

## Reporting policy inconsistencies and issues

Other participants reflected on the reasons for reporting not extending far enough to capture what they deemed to be offensive or abusive content:

***I'm thinking particularly of Facebook here, cause that's where I've encountered the most content, but I think they could do with broadening their views on what is offensive content***

(FG15: LGBTQI+, 22-39 years, female).

This was a common view held by participants and one that significantly impacted their confidence in social media services and online dating apps' commitment to respond to reports of offensive content, to provide a safe environment for all users, and to have policies in place that reflect social norms. As these participants observed:

***I would report it, but I would also have absolutely no faith that it would actually do anything because every time I've ever reported anything like that [racist posts], it just automatically comes back as not a violation***
(FG5: 26-52 years, female).

***Most of the time I've seen the stuff that is harmful to me, and I try to report it on Instagram, and a couple of days later I get a message back, "It was not a valid reason to report on, and we're not going further with it"***
(FG3: 28-51 years, female).

***I've had times, "We've reviewed it and it doesn't go against community guidelines", but to me, it very clearly does if it's racism or hate speech, and things like that***
(FG3: 28-51 years, female).

Participants identified these concerns in response to a range of factors, but they were consistently noted by participants in the context of racist posts and anti-LGBTQI+ posts:

***I absolutely like certainly [have] seen some really horrific kind of racial stuff that's allowed to stay up and like a really awful violent speech against trans people. And that's been reported … I've either done it myself or know that others have reported it and it's, the platform literally replies and says, "no, we've reviewed this and it's fine"***
(FG15: LGBTQI+, 22-39 years, female).

Another participant observed:

***I lodged a complaint probably a couple of months ago about what I thought was racist comments. I'm an Indigenous man by the way, and I thought they were racist and offensive. … I'm not particularly thin-skinned about this sort of stuff. I mean, some people can be very thin-skinned about it. I'm not. But I found it pretty offensive. … And I'm still none the wiser as to what the measure was. They said, "no, it wasn't racist". And I'm still unaware as to what the basis of their decision was***
(FG19: Aboriginal &/or Torres Strait Islander, 20-68 years, male).

In the context of her experience as an Aboriginal woman, another participant remarked:

***I would like these companies to understand what it's like to be a mum or a woman who's constantly hypervigilant, and have that awareness, I suppose***
(FG1: 25-49 years, female).

One participant described the approach used by social media services and online dating apps as:

***Sort of skewed, unfortunately to the detriment of certain identities***
(FG15: LGBTQI+, 22-39 years, female).

In expanding on this view, the participant described how policies appear inconsistent in how they respond to reports of offensive content, whereby abusive and harmful comments and posts are not considered to violate platform policies. However, general lifestyle posts from people in specific communities, such as the Queer community, have non-offensive content removed. They observed:

***There seems to be a lot [of] double standards. … I follow a number of Queer users and their accounts are very benign. Nothing saucy, just living their lives and they've had things taken down because of their sexuality, even though they're not talking about their sexuality, you know. In that case, it seems almost skewed the wrong way I suppose. … Like a little bit sort of homophobic in its actual policing of what's offensive. … I feel like none of the platforms I use might have that right***
(FG15: LGBTQI+, 22-39 years, female).

Another participant also described this inconsistency in defining offensive or inappropriate content in platform practices in relation to moral judgments made around consensual photos with partial nudity, versus statements that are likely to fuel racist attitudes. This participant described reporting content they viewed as somewhat akin to hate speech which was ***"spreading ideas of different groups which aren't true"*** (FG3: 28-51 years, female). They used an example of a report they made about a post downplaying the murder of

Aboriginal people during colonisation, versus a situation where their friends post partially nude (consensual) images of themselves at the beach:

***I reported it. …That [report] came back and said it was in line with community guidelines, but then a lot of friends would post bikini pictures … and then that's taken down within five hours because it's so against community guidelines. So, to me, there's just that disparity of what's considered against community guidelines***

(FG3: 28-51 years, female).

Others similarly described how they found inconsistent responses depending on how they reported content. As one participant observed, if offensive content was reported as spam, social media services and online dating apps would remove it. But if the same content was reported as harassment, it was deemed not offensive. They explained:

***If I report something as spam, [it] will get actioned. If I report something as harassment, bullying, hate speech, [it] doesn't go through, it doesn't get actioned. … Now if I see something as bullying, I would report it as spam and it gets more looked at, which is a little bit crazy. … So, it's better to report everything as spam***

(FG14: 18-40 years, female).

Participants also reflected on how the inconsistency across social media services and online dating apps regarding what constitutes offensive content created difficulties in staying safe online. As one participant noted:

***What you say on one platform might not be considered hate speech, for want of a better word, and on another platform it is. So, wouldn't it be nice if there was some form of consistency across everything***

(FG3: 28-51 years, female).

Young people in the 13-15 year focus groups reported that it was relatively simple to play with or game filters and other control features by banning 'mild' words, using ordinary terms or euphemisms as slurs or insults (e.g., I don't want a bunch of angry comments from bri*ish people) and finding ways around filters (e.g., using pr0n, p*rn or p0rn). As one participant explained:

***There's always people who could find a way to get around posting that content and you'll probably see it sometime***

(FG21: 13-15 years, mixed).

This ability to 'game' the functions was considered among these groups to reduce the effectiveness of user control features.

## Lack of 'human' moderators and too much reliance on automated detection and responses

Further to inaction and inconsistency in response to reports, one of the common concerns raised by focus group participants relating to the ineffectiveness of user control features was the absence of knowledge or follow-up once something was reported to a platform. There was a sense that most reports go through some form of AI process, in which they are assessed by a machine, as opposed to a human moderator:

***The thing that annoys me the most is that for the vast majority, it's all automated, it's [a] computer that actually do[es] the checking and not real people. And that's why it's not [working] properly***

(FG14: 18-40 years, female).

***It's almost never a human being actually figuring out what happened and whether there's actually been something wrong***

(FG5: 26-52 years, female).

This issue raised several concerns for participants in assessing the effectiveness of user control features, notably that it showed a lack of care or commitment from the social media services and online dating apps to keep users safe and make them feel heard. As one participant observed:

> ***When they make it very difficult for you to find the ways to report something, or you get an automated message when you report something as being x, y and z, and you get a thing saying, "Oh actually, it's not against our guidelines so we're going to keep it up", I think that makes it feel like your safety is not really being taken into consideration***
>
> (FG5: 26-52 years, female).

Other participants similarly reflected on the lack of a human moderator or contact point as making them feel the social media services and online dating apps ***"just don't care at all"*** (FG5: 26-52 years, female):

> ***Facebook just doesn't care about [users]. … I don't even think about actually reporting content anymore because I just know that Facebook and TikTok are not going to do anything***
>
> (FG5: 26-52 years, female).

Another participant similarly expressed this view, arguing that there needs to be ***"humanity brought back into it"*** (FG9: 32-39 years, female). They explained:

> ***There's no actual person that's dealing with this. So, it's all computer botted, it's all automated, there's no personalisation in the fact that something has happened to someone and there's actually a human who's involved and is going to refer this or get involved. It's like if it doesn't fit the scenario, we can't do anything about it. It doesn't matter how you feel about it, we won't act. So, I think there has to be some sort of personalisation and humanity brought back into it because we're humans, we all have different reactions to things, and so nothing just always fits in the box***
>
> (FG9: 32-39 years, female).

## No updates or delays on the outcomes of reports

Participants also commonly expressed experiencing a lack of follow-up when they did report someone, which contributed to a feeling of ineffectiveness. As this participant described:

> ***I have had no satisfaction with Facebook in terms of trying to report anything. … Every single thing that I've tried to, I suppose report to them in terms of just even for the greater good, as I said, a serial pest more or less, not once have I ever had a response of, you know, X, Y, Z has now happened***
>
> (FG18: 23-47 years, female).

Participants felt there should be more onus on social media services and online dating apps to provide an update on the outcome of a report and encourage educational or safety messages for users, so they can find out where to access support beyond the platform if needed. As this discussion indicated:

> ***P1: It just feels like if I report something, it feels like it's sent to the abyss, there's no one there … and there's nothing there to reply back to me about that. That's what I feel like. So, I feel like a presence would be nice from these people who run them.***
>
> ***P2: Yeah, or it could be – you just reminded me, say when you report on something and you get the message saying, "Thanks for reporting this", whatever, [if] it had "And here are the local resources. This is the contact for the eSafety Commissioner. If you feel like you're unsafe, you can call 000". That would be really helpful. Location specific information or resources around what else you can do outside of within the app. I think that would encourage me more to take things out of the app than just sit and wait for a miscellaneous response***
>
> (FG5: 26-52 years, female).

Other concerns around the effectiveness of reporting to social media services and online dating apps were that even when the content was deemed offensive or inappropriate, it took a long time for anything to be actioned or content to be removed. As these participants explained:

***P1: A fake account was made like impersonating her [a friend], implying that there would be a lot of sexual content on it. So being like, "follow me for explicit content" and using the pictures of her and it took months for it to be shut down and you know, it was really distressing for her in the meantime to have this account in her name basically set up.***

***P2: We went through the process of trying to get [a fake profile] shut down … and that was going on for about two months before Facebook was like, oh okay, yeah, it's a scam, we'll shut it down. So that was pretty frustrating***

(FG18: 23-47 years, female).

## Unintended consequences of control features

Some research participants identified that there can be trade-offs or unintended consequences resulting from the introduction of control features developed to improve safety. Some participants commented on a control feature to block contacts on Tinder – designed to avoid you connecting with an ex-partner, or with people you may not want to know you are on the app, such as a work colleague or friends – having the inadvertent unintended effect consequence of allowing people to cheat on their partners:

***What if someone utilises that and they cheat on their boyfriend or girlfriend?***

(FG7: 25-54 years, mixed).

***I think that feature can actually perpetrate cheating***

(FG17: 44-69 years, female).

Another example of an unintended consequence raised in response to Grindr's policy to censor racial preferences on profiles was that it has reduced opportunities to visibilise and challenge racial discrimination. As one participant commented:

***Sometimes I prefer that people be honest and not kind of walk around difficult conversations and if we just ignore things then no one talks about it or just tries to be really, you know, hyper progressive and not say things. And not be able to learn. So even though those things are not nice to see, I also like to see if that's the reason people don't wanna talk to me, you know, I'd like to know that instead of, you know, not knowing***

(FG16: LGBTQI+, 27-30 years, male).

This sort of unintended consequence requires careful consideration by social media services and online dating apps and ideally engagement with relevant users when developing control features.

## Lack of shared information across social media services and online dating apps and ineffective user identification confirmation processes

Another common criticism of the effectiveness of platform control features was that abusive people's information is not shared across social media services and online dating apps. As one participant observed:

***Collecting behaviour [on people]. … You know, he's a sex pest. I want to know about that or avoid that if I can. … I think the follow-up [by platforms] would be good to know, ok, maybe this person has done this multiple times. … If someone keeps doing something, something needs to happen***

(FG13: LGBTQI+, 18-47 years, mixed).

Another participant similarly claimed:

***Platforms should be doing regular account verifying and checking to see if that account should be allowed on their platform in the first place. Just like [a] history of what they've been posting, or things that they've been liking, might give them an idea of, "Okay, this person shouldn't be allowed"***

(FG2: LGBTQI+, 17-24 years, female and non-binary).

This was a prominent concern raised regarding online dating apps, whereby someone blocked or reported on Tinder could still be active on Bumble. But it also came up in relation to blocked users simply creating new accounts and being able to re-approach or harass people after they had already been blocked or banned:

***I know on Grindr you get quite a few pest people that just will message you and then you block them and then just like they keep creating new accounts and they just keep doing that***

(FG16: LGBTQI+, 27-30 years, male).

***When you block a contact and before you know it, a contact, this person, creates another account and it appears back to your contact list. It [is] kind of frustrating***

(FG15: LGBTQI+, 22-39 years, female).

Another participant experienced a similar situation and described it as follows:

***I had someone who matched, and they were just a nutcase and so I deleted them, unmatched [them]. Then I had another one that was saying exactly the same thing and it was them because I could tell, they were using the same words, the same sentences, different photos, and then they would always say the same things and it would always end the same way. And no matter what you did, you couldn't really get rid of them because it's a new account or they've got [a] multitude of phone numbers, because you don't even need to have a phone number anymore to register, you can do it by email***

(FG9: 32-39 years, female).

This participant flagged that this safety concern has been exacerbated with the linking of social media services and online dating apps, which makes it easier for people to track users across social media services and online dating apps. They explained:

***With Tinder now … because it uses Facebook, those people will start coming up on your "people you may know" lists with Tinder, and that's a bit confronting because you can block them on Tinder, but they'll still come up on your Facebook and probably see more about you***

(FG9: 32-39 years, female).

One common suggestion to improve user safety on social media services and online dating apps and respond to the concerns mentioned above was the need for social media services and online dating apps to enhance the user identity confirmation process:

***So, I'm thinking along that wavelength if, especially the dating apps, not so much Facebook and Instagram, but to actually legitimise a person is actually the person who is on that profile. … So, when you are basically filtering through people, you can see the legitimate ones compared to the ones that haven't actually been verified and then the user can make the decision from there***

(FG14: 18-40 years, female).

Others suggested ***"doing background checks on the individual"*** (FG14: 18-40 years, female):

***If you had to have your own real life details on, on your online accounts, then I think that would do away with a lot of the harm that is being done***

(FG13: LGBTQI+, 18-47 years, mixed).

***Before you want to get on the app and make up all this garbage about yourself, like after getting a background check, like having a score or something before you're allowed on the app. Cause maybe that would entice me to feel more comfortable on that dating app knowing that the people that go on it have to have a certain credit score or I know that they've gone through a process already, like a background check***

(FG14: 18-40 years, female).

Enhanced user identity confirmation processes were also raised concerning other social media services, beyond dating apps, with several participants reflecting on their experiences either with others or themselves creating multiple accounts using the same email or using numerous different emails to create multiple accounts:

***You should only be able to make one account linked with the email, because … [I have] I think up to five accounts with the same email, the same password, and you just make your username and people can catfish and spam and whatnot on the other accounts***

(FG24: 13-15 years, mixed).

Another participant explained:

***You don't actually need to prove yourself to create an account. Sometimes you would see the same person or the same group trying to do the same thing, or crime, on social media, which is really annoying. … Like even us as a user we could easily create 10 different accounts under different names***

(FG5: 26-52 years, female).

Other participants suggested more effective verification processes should be implemented by social media services and online dating apps. As one participant stated:

***With Instagram and Facebook, there's no facial recognition, unlike Bumble, so I feel like anything, and anyone, can use Instagram and Facebook.***

(FG5: 25-52 years, female).

Participants likened improved identification standards with common offline practices, for example, being part of a reading group for children, library membership, and drawing on historical examples, video store membership (where people could borrow videocassettes, DVDs and blue rays to watch at home and later return). As these participants from the same focus group explained:

***I go to the school to help with reading groups, and I had to have so many checks. I had to bring in my driver's license, my Medicare card, they had to make sure I was who I was saying. But on the platforms, I can go on there and just say, "oh I'm [Name], I'm this and this", and there's no consequences. I can just lie. I can do whatever I want.***

…

***I do remember back in the day, even to join a video store, the amount of ID [needed]. I had to show like the driver's license just for that. But now that these platforms where there's more chance of bullying happening or being contacted by someone who is a threat to you, I mean all they need to join is a mobile number and an email, which people can hide behind and there's no photo ID***

(FG17: 44-69 years, female).

Encouraging social media services and online dating apps to implement more stringent identification standards was supported by many participants as a way to enhance user safety online, as reflected in the following comment:

***I like the idea of having profiles ID verified maybe like a photo with your driver's license and your face from today so they can verify that that's what you look like and that's your ID***

(FG14: 18-40 years, female).

# Matching Australian users' expectations and knowledge with what social media services and online dating apps purport to do

We provided respondents with a selection of scenarios to ascertain their knowledge of social media services and online dating apps' control features, their view of the effectiveness of these control features, and to gain some insights into how their experiences and understandings aligned with what each service and platform has publicly purported that it can do. Each focus group was presented with two scenarios from five options, except the 13-15 year focus groups who, given their age, were only presented with the social media service options (not the online dating app scenarios). The control features are detailed in Table 2 below. The focus group facilitators selected the scenario/s that were most relevant to the research participants, based on the social media services and online dating apps they identified as using (this was the first question of the focus groups).

### Table 2. User control safety feature scenarios

| Platform | Safety Feature |
|---|---|
| Bumble | Private detector tool |
| TikTok | Blocking potentially offensive content |
| Tinder | Blocking contacts |
| Meta / NCII | Digital hash for intimate images |
| Instagram | Filter racist posts* (note: this has since been expanded to all hateful messages, not only those containing racist content) |

Participants were then asked the following questions:

- Are you aware of this feature?
- Would you use it or recommend its use to others? (Why/Why not?)
- Do you believe it would be effective? (Thinking back to the scenarios, would this feature address some of the issues raised? Or are there any ways you think this feature could be improved to address online safety?)

As indicated by the comments below, very few focus group participants knew of the user features available on the social media services and online dating apps, despite these being promoted on the platforms and despite respondents regularly using these social media services and online dating apps:

***I haven't seen that, but I think that's a good idea***

(FG4: LGBTQI+, 27-44 years, mixed).

***I don't think I've heard of that function***

(FG5: 26-52 years, female).

In discussing these control features, many respondents supported the initiatives. Support was particularly strong for features that helped prevent potentially abusive content from being received by the user in their Instagram direct messages:

***I think it's a good idea as well. … It's good for people that get offended from racist posts***

(FG13: LBTQI+, 18-47 years, mixed).

However, participants did not feel the racist post feature went far enough and argued that it should be made available in contexts beyond direct messaging and racist posts. For example, to include homophobic or misogynistic content or other language/emojis/content that may be offensive to an individual:

***Filter racist posts has got potential. I was going to say the same thing. I'm not sure why it's filter racist. And I wasn't clear whether you can set yourself the words that you want to – because if you can set the words yourself, then it doesn't – it's not really just racist, is it? But I think that's got potential. … I can see the good use for that; to protect yourself from being called those names***

(FG1: 25-49 years, female).

***A broader feature would be better. … A personalised version. So, let's say, I didn't want any dating advances on Instagram, I could choose that***

(FG9: 32-39 years, female).

***If it was customisable. … I personally haven't had any real instances involving racism or abuse in that sense, so it wouldn't be as useful to me. But I'm sure it could be if it was developed more and in a wider range of situations and instances***

(FG9: 32-39 years, female).

Participants also felt the feature would be useful across all social media services and online dating apps, not just Instagram:

***It could apply [to] all the different social media, couldn't it – all the different platforms, not just Instagram. … I will definitely use them if available on Bumble … or even other platforms***

(FG6: 55-66 years, female).

Participants were particularly supportive of a feature that gave them the control to determine what they found offensive, as opposed to an automated feature that predetermined what was offensive:

***I think it could be a really useful feature if you can figure out like which terms or tags are most associated with the content you don't want to see, then you can add that … and then anything that features that tag won't show up in your feed, which would be nice***

(FG4: LGBTQI+, 27-44 years, mixed).

***I feel like it's good. … I don't want to hear about men's rights movements or #vaccineconspiracy or something, so I'm going to block that content***

(FG1: 25-49 years, female).

While recognising the need for user controls in this instance, for some of the other features, respondents felt the social media services and online dating apps should have an 'opt-out' instead of an 'opt-in' mechanism. For example, it should automatically have users 'opted-in' to privacy settings and blocking violent or harassing content, rather than the user having to locate and find how to implement these control features. As these participants explained:

***I like the idea of using those AI features as well to censor information, and for users to opt-in to see them rather than them just appearing there***

(FG3: 28-51 years, female).

***Going back to what you were talking about earlier about additions to the platforms themselves, with the stuff like the filter racist posts tool … even if you had it turned off … it might be good to have a notification or something to suggest that you do turn it on if you want it, so it's more accessible instead of having to go into a settings menu***

(FG21: 13-15 years, mixed).

## Age appropriate engagement

While control features related to age were not specifically presented to participants in the scenario discussions, this was a common theme that emerged across all participant groups. Participants in the 13-15 year focus groups described the need to protect children from non-age appropriate content and online predators. As one participant explained:

***Little kids and stuff might talk to someone, and they might say, "Oh hang out with me. I'm your age" or whatever and then it can – I've seen stuff like people – like little kids doing stuff that they shouldn't, because older people posing as younger kids ask them to do stuff***

(FG24: 13-15 years, mixed).

Another participant similarly described observing adults trying to connect with young people online:

***People who like are way older that lie about their age to get close to younger people, because you don't always know, especially on sites where you can't always necessarily send photos and stuff***

(FG24: 13-15 years, mixed).

This view expanded beyond the 13-15 year focus groups to other groups as well:

***It's very easy these days I feel like for young people to be groomed by older individuals, whether it be through coercion or talking to an individual until they become older. Because what can happen is, if an older individual has really easy access to a younger individual, what can happen is, it's a very big security risk for the younger individual … because what happens if this online relationship goes off a platform into in-person?***

(FG2: LGBTQI+ 17-24 years, female and non-binary).

Participants felt that social media services should be responsible for ensuring age appropriate content and control features, and enhancing confirmation processes for checking users' ages to better align with what social media services purport to do, versus what happens in reality. As these participants observed:

***I agree with [another participant's] idea to make some way to verify your age, because it's kind of a problem on TikTok. I get so many videos of little kids, actual kids posting videos and then people in the comments being mean to them***

(FG21: 13-15 years, mixed).

***Sometimes it's kids saying they're older than they actually are on the app, so maybe they should make them verify their age to make sure people are safe online***

(FG21: 13-15 years, mixed).

***I think the most important thing is because it's from personal experience, I think platforms really need to make sure about the ages of their user base***

(FG2: LGBTQI+ 17-24 years, female and non-binary).

Other suggestions to address issues regarding age were to create ***"a kids' version of something so that kids can enjoy it and it's safer for kids rather than older people who use it"*** (FG21: 13-15 years, mixed) and to put the onus on social media services to engage with young people:

***These platforms need to make younger users more aware of their own safety and the responsibilities. They [the platforms] need to have more open support networks and to make those support networks more known [to younger users]***

(FG2: LGBTQI+ 17-24 years, female and non-binary).

## Education

As discussed earlier in this Report, a lack of awareness of control features available on social media services and online dating apps was common:

***There should be more awareness about the different features and settings and stuff. … It's hard to know what they [control features] mean. … Maybe the app could post a notification or something?***

(FG23: 13-15 years, mixed).

Education was seen as critical to bridging the gap identified in the focus groups between user experiences and awareness of control features and what is publicly purported on social media services and online dating apps. As one participant explained:

***The fact that it [the safety feature] exists is really good. But what needs to be done, it needs to be made more aware to people. Because the problem is, support is only effective if you also know about it***

(FG2: LGBTQI+, 17-24 years, female and non-binary).

Who participants felt should be targeted in relation to enhanced education ranged from young people in school settings, to more frequent messaging to platform users when they sign in or post on social media services and online dating apps, to more broad messaging and education for the general public, for example, campaigns and advertising:

***More education and more promotion needs to be done in the primary and for the higher school students, because even though you give them a lot of information, sometimes they get impulsive***

(FG3: 28-51 years, female).

***There needs to be more discussion in the school, because I know only once a year the police, or somebody, comes to discuss all that stuff with the kids, and for the entire year the kids are with their own technological devices***

(FG3: 28-51 years, female).

***I have seen a couple of things about digital footprint … [and] it's not actually said how much it can actually affect you, because at my school, they said, "Digital footprint", and then you don't actually think it's a real thing until it actually happens to you. So, I feel like they should actually explain how it can affect you in the long term because people have lost their jobs from it and stuff***

(FG21: 13-15 years, mixed).

***Actually, something that could be done more for dating apps, but it could be done for everything, … you have to go through a training on what's acceptable, [what] wasn't acceptable and you know, people that are really bad it won't change much but I don't know it can. Some people are just unaware, like some people think it's okay to, that "no means yes", and things like that***

(FG14: 18-40 years, female).

Participants also pointed to the importance of educating users who are unsafe online or who engage in abusive or inappropriate conduct. As one participant described:

***I think that's a missing part. Educating offenders. I know, like when I've made mistakes in the past just to find out what you can do or where you can learn things that's not very accessible for offenders. But if you are a victim, all the information you know is there. But on the opposing side, you don't really know where you're going wrong at times***

(FG16: LGBTQI+, 27-30 years, male).

Ultimately, participants felt there was very little difference between the potential harms experienced online versus offline. However, participants expressed that there should be a greater onus on social media services and online dating apps to reflect social norms and standards and hold people to account, in the way people are held to account offline (e.g., by police and laws). Participants also felt more accountability should be placed on social media services and online dating apps to improve their control features, improve their education and awareness raising of these features for users, and provide more effective policies and responses. As these participants stated:

***I guess it's not so much about what I want apps to know about online safety, but how I would like their safety features to make me feel, is to make me feel like I'm being listened to and that my safety is being taken seriously, no matter how I define what is safe or unsafe and what isn't***

(FG5: 26-52 years, female).

***I just want them to make us safer and, you know, protect us. … The internet has been around [for] awhile now, but it just seems so weird that in real life there's all these laws and everything, but then online it's different. Like online you can be harassed, and I don't really know that there's that much you can do about it. So, I guess I would like online for you to be as safe as you are when you're in real life. If someone does the wrong thing, it should be the same. You can, if a law's broken, that you can report things and that there's consequences***

(FG4: LGBTQI+, 27-44 years, mixed).

# Implications for industry

Participants held positive views of some social media services and online dating app control features but were unaware of others. The effective features participants identified included the ability to report fake profiles, automated control features that prevent future profiles with the same contact information from following or friending users, AI functionality that blurs potentially offensive images, the ability to block contacts, and to filter out words or hashtags they do not want to be exposed to. Community notes, a crowd-sourced fact-checking program, was also viewed as productively adding context to potentially harmful tweets. These features can help users feel safer and more comfortable using social media services and online dating apps and indicate that a platform is taking their safety seriously. Users who identified effective mechanisms also expressed that they were more likely to continue using the social media service and/or online dating app and recommend it to others, than those who reported negative experiences with control features. This suggests that it is in the interests of social media services and online dating apps to invest in developing and implementing effective user control features, policies, and automated control features, as they could improve the user experience and help to address some of the negative aspects associated with social media use, such as cyberbullying and harassment.

On the shortcomings of control features, participants noted that users employ various strategies (such as creating new accounts) to bypass mechanisms and sanctions. Younger participants in the 13-15 year focus groups in particular, spoke of ways to circumvent filters. For these reasons, and despite being familiar with and having deployed the reporting and blocking functions, some participants did not believe these enhanced or protected their safety.

## Improve reporting

Those who were more critical and dissatisfied with control features, particularly the reporting functions, described social media services and online dating app responses to user reports as inconsistent, delayed, or non-existent. Some respondents believed that reporting categories need to be expanded. Many expressed confusion over what constituted offensive content, particularly when what might be considered acceptable on one platform was not acceptable on another. The lack of responsiveness, transparency and consistency in reporting decisions was considered to result in frustration and a lack of trust in social media services and online dating apps' systems. This was identified as a particular concern for vulnerable and minoritised communities by participants who were representative of these groups, and those who were not. **Participants urged social media services and online dating apps to ensure that their policies are consistent, transparent and regularly reviewed.** They suggested that social media services and online dating apps should provide updates on the outcome of a report and offer educational or safety messages to assist users in accessing support beyond the platform.

> **Users who identified effective mechanisms also expressed that they were more likely to continue using the social media service and/or online dating app and recommend it to others, than those who reported negative experiences with control features.**

## Develop control features for all, especially those most at risk of harm

Some participants mentioned that they did not use various user control features due to the belief that such features were not created for them. This sentiment was particularly reflected amongst participants experiencing, or at risk of, multiple forms of discrimination and disadvantage. As discussed earlier in this Report, there was a clear sense that inconsistent standards existed whereby vulnerable and minoritised identities were more at risk of being reported and having their content removed, than having their concerns of harmful content or abuse being listened to. This is reflected in academic research which has similarly found that vulnerable and minoritised communities' needs and experiences are not prioritised by social media services (Carlson, 2020; Kennedy, 2020; Matamoros-Fernández, 2017). As a result, users may rely more on self-created filters or take privacy and safety management into their own hands due to a lack of trust in the social media services and online dating apps' ability to protect them. **These findings suggest there is a need for social media services and online dating apps to consider the diverse needs of user groups in designing control features and policies.**

> **Participants called for platform policies and regulation practices to reflect social norms and values and the needs and lived realities of diverse, vulnerable and minoritised Australians.**

Questions about the ability of social media services and online dating apps to protect these communities were also raised in relation to how they define and respond to offensive and abusive content. Participants called for platform policies and regulation practices to reflect social norms and values and the needs and lived realities of diverse, vulnerable and minoritised Australians. Ultimately, participants called for social media services and online dating apps to adopt human-centred approaches which reflect the diverse and complex needs and identities of all users, including through individualised, customisable control features. Social media services and online dating apps must be responsive to minoritised and vulnerable users. Addressing this does not require a reduction in the scope of rights, but rather an increase in functions tailored to meet the needs of those at greater risk.

## Continue to improve workforce diversity

Social media services – such as Meta – have made attempts to address the underrepresentation of women and gender and sexually diverse people, Indigenous and First Nations people, culturally and linguistically diverse people and people with disabilities, in development and design roles in the tech industry (Meta, 2022). This is significant, as scholars maintain a lack of diversity contributes to the high rates of victimisation to which these cohorts are subjected, as their experiences, needs and risk may not be anticipated or addressed by those in positions of power (Chang, 2018; Harris & Vitis, 2020; Ionescu, 2012; Reed et al., 2018; Suzor, 2019). This can, as participants highlighted, result in persons from minoritised communities experiencing higher levels of online harm and greater efforts to protect themselves online.

## Improve usability and awareness of control features

Overall, the focus group discussions revealed a lack of awareness among users about the different control features available. Some were unfamiliar with how to access and use control features and flagged that there was a lack of information about navigating them. This suggests that there is a need to improve the ease of use of control features. When participants were presented with control features available on various social media services and online dating apps, very few participants knew about them, despite being frequent users of those platforms. Many participants supported these control features or recommended the creation of features, which unbeknown to them were already in existence, especially in preventing abusive content in direct messages. However, they felt that they should be customisable to filter out other offensive content beyond racist commentary, expand beyond direct messages to all content received, and extend to all social media services and online dating apps.

## Inform and empower users

Participants preferred controlling what they found offensive rather than having this determined by an automatic feature. Some participants suggested having an 'opt-out' mechanism for certain control features (that were automatically enabled) rather than having to locate and implement them. Users recommended that changes in functionality and features that occur with updates should be clearly communicated to users.

**Participants suggested more education on user control features, including increased education among young people in schools, more frequent education messaging to platform users, and widespread awareness campaigns and promotion of safety mechanisms for the general public.**

Deficits in participant knowledge and confidence in managing their use of online dating apps and social media services reinforced the need for information and awareness-raising campaigns. Many participants reported discovering control features accidentally or through word-of-mouth from friends. Social media services and online dating apps could address this knowledge gap through periodic, pop-up reminders of control features, in the way that some dating apps, for instance, prompt and remind users about paid versions of these apps (like Grindr Xtra or Tinder Premium). For instance, Facebook periodically reminds users to check their privacy settings through a message in the newsfeed, and this nudging could be replicated in other ways and on various social media services and online dating apps.

Some younger participants learned about online safety through the school curriculum and police visits to their schools. A few participants (mainly younger Australians who had completed cyber safety programs at school) mentioned the eSafety Commissioner as a helpful resource. However, many participants lacked knowledge about third-party solutions and online safety laws. **These findings suggest that organisations like the eSafety Commissioner can play a significant role in sharing information about online safety, particularly because participants who had knowledge of information and resources provided by the eSafety Commissioner** described this as a very useful resource. Participants felt it would be helpful to have more accessible easy-to-understand information and greater confidence about third-party solutions and laws. However, they emphasised that the social media services and online dating apps themselves are responsible for improving user awareness of user control features. Participants also underlined the importance of promoting behaviour change in users who engage in abusive or inappropriate conduct through education and awareness raising of the harms of their behaviour.

## Human moderation

Many participants believe that social media services rely too heavily on automated processes to assess reports about harmful content and behaviours online, instead of involving human moderators. This made them feel as though their complaints and requests were unheard and that they were unsafe. Participants reported feeling annoyed and dissatisfied when receiving automated responses, which they felt showed a lack of consideration for their concerns and well-being. They saw potential in automatically filtering out blatantly offensive content, but called for a more personalised and human-driven approach to dealing with harmful content and behaviour on social media services. Participants noted that AI regulation was unlikely to understand comments, especially involving diverse communities and identities.

> **Participants noted that AI regulation was unlikely to understand comments, especially involving diverse communities and identities.**

## Improve online safety for children

Another primary concern that participants had was the safety of children online. They called for social media services to do more to confirm the ages of users and ensure age appropriate content on digital media and control features are offered. Participants also suggested putting more onus on social media services to engage with young people and make them aware of control features. These concerns were expressed across all participant groups, particularly among the 13-15 year focus groups. Improved identification confirmation processes by social media services were also proposed. Participants supported more stringent standards that ensured social media services could accurately confirm users' identities.

# Conclusion

Our findings suggest that social media services and online dating apps should invest in developing and implementing effective user control functions and automated control features that are accessible and useable for all, including vulnerable and minoritised users. Additionally, the study highlights the need for improved education and awareness campaigns about control features, safety mechanisms and industry and government regulation. As our walkthrough analysis revealed, many social media services and online dating apps have nuanced user control functions, but our participants were often unaware of these features, beyond blocking and reporting. Improving education and awareness has the potential to ensure that Australians better understand online safety and available resources and to bolster user experiences online.

Participants felt that the onus should be on social media services and online dating apps to improve and ensure user safety through user control features, digital platform policies and moderation, and to be transparent about regulation processes. They believed that there could be more effective and proactive responses to potential online harms and that social media services and online dating apps should hold users accountable for their behaviour online and seek to prompt behaviour change. As outlined above, information-sharing, education and awareness raising channels and campaigns were recommended to aid this process.

**Australians of all identities want the option and freedom to use online digital dating apps and social media services and to have their safety prioritised online, as it should be offline. The role and uptake of technologies speaks to the urgency in strengthening and extending policy and practice to accomplish this, and the potential benefits to the lives, wellbeing and engagement of all Australians.**

**Improving education and awareness has the potential to ensure that Australians better understand online safety and available resources and to bolster user experiences online.**

# References

AppMagic.Rocks. (2023). *TopApps*. https://appmagic.rocks/top-charts/apps?tag=103&date=2023-01-01&country=AU

Australian Bureau of Statistics [ABS]. (2016). *Census of Population and Housing: Index of Relative Socio-economic Advantage and Disadvantage (IRSAD) Socio-Economic Indexes for Areas (SEIFA), Australia, 2016.* https://www.abs.gov.au/ausstats/abs@.nsf/mf/2033.0.55.001

Brown, C., Sanci, L., & Hegarty, K. (2021). Technology-facilitated abuse in relationships: Victimisation patterns and impact in young people. *Computers in Human Behavior*, 124. https://doi.org/10.1016/j.chb.2021.106897

Byron, P., Robards, B., Hanckel, B., Vivienne, S. & Churchill, B. (2019). "Hey, I'm having these experiences": Tumblr use and young people's queer (dis)connections. *International Journal of Communication*, 13, 2239-2259, https://ijoc.org/index.php/ijoc/article/view/9677

Cama, E. (2021). Understanding experiences of sexual harms facilitated through dating and hook up apps among women and girls. In J. Bailey, A. Flynn, & N. Henry (Eds.), *The Emerald international handbook of technology-facilitated violence and abuse.* (333–350). Emerald Publishing. https://doi.org/10.1108/978-1-83982-848-520211025

Carlson, B. (2020). Love and hate at the Cultural Interface: Indigenous Australians and dating apps. Journal of Sociology, 56(2), 133–150. https://doi.org/10.1177/1440783319833181

Carlson, B. (2021). Why social media platforms banning Trump won't stop—Or even slow down—His cause. *The Conversation.* http://theconversation.com/why-social-media-platforms-banning-trump-wont-stop-or-even-slow-down-his-cause-152970

Carlson, B., & Frazer, R. (2021). Indigenous digital life: The practice and politics of being Indigenous on social media. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-84796-8

Carlson, B., & Frazer, R. (2018). Cyberbullying and Indigenous Australians: A review of the literature. Aboriginal Health and Medical Research Council of New South Wales and Macquarie University. https://research-management.mq.edu.au/ws/portalfiles/portal/92634728/MQU_Cyberbullying_Report_Carlson_Frazer.pdf

Chang, E. (2018) *Brotopia: Breaking Up the Boys Club of Silicon Valley.* Portfolio/Penguin Books.

Department of Social Services. (2022). National Plan to End Violence against Women and Children 2022-2032: Our commitment to ending all forms of gender-based violence [Ten Year Plan]. Commonwealth of Australia. https://engage.dss.gov.au/wp-content/uploads/2022/01/Draft-National-Plan-to-End-Violence-against-Women-and-Children-2022-32.pdf

Duguay, S. (2020). You can't use this app for that: Exploring off-label use through an investigation of Tinder. *The information society*, 36(1), 30-42. https://doi.org/10.1080/01972243.2019.1685036

Duncan, Z., & March, E. (2019). Using Tinder® to start a fire: Predicting antisocial use of Tinder® with gender and the Dark Tetrad. *Personality and Individual Differences*, 145, 9–14. https://doi.org/10.1016/j.paid.2019.03.014

Flynn, A., Hindes, A., & Powell, A. (2022). *Technology-Facilitated Abuse: Interviews with victims and survivors and perpetrators* (Research Report 11/2021). ANROWS. https://research.monash.edu/en/publications/technology-facilitated-abuse-interviews-with-victims-and-survivor

Gillett, R. (2021). "This is not a nice safe space": Investigating women's safety work on Tinder. *Feminist Media Studies.* https://doi.org/10.1080/14680777.2021.1948884

Harris, B., & Vitis, L. (2020). Digital intrusions: Technology, spatiality and violence against women. Journal of Gender-Based Violence, 4(3), 325–341. https://doi.org/10.1332/239868020X15986402363663

Harris, B., & Woodlock, D. (2021). *'For my safety': Experiences of technology-facilitated abuse among women with intellectual disability or cognitive disability.* The eSafety Commissioner. https://www.esafety.gov.au/sites/default/files/2021-09/TFA%20WWICD_accessible.pdf

Henry, N., Vasil, S., Flynn, A., Kellard, K., and Mortreux, C. (2022). Technology-facilitated domestic violence against immigrant and refugee women: A qualitative study. *Journal of Interpersonal Violence*, 37/13-14: 12634—12660. https://doi.org/10.1177/08862605211001465

Ionescu, A.C. (2012) ICTs and gender-based rights. *International Journal of Information Communication Technologies and Human Development*, 4(2): 33–49. http://doi.org/10.4018/jicthd.2012040103

Kemp, S. (2023). *Digital 2023: Australia. We are social and Meltwater*. https://datareportal.com/reports/digital-2023-australia

Kennedy, D. T. (2020). *Indigenous Peoples' Experiences of Harmful Content on Social Media.* Macquarie University. https://research-management.mq.edu.au/ws/portafiles/portal/135775224/ MQU_HarmfulContentonSocialMedia_report_201202.pdf

Light, B., Burgess, J., & Duguay, S. (2018). The walkthrough method: An approach to the study of apps. *New Media & Society*, 20(3), 881–900. https://doi.org/10.1177/1461444816675438

March, E., Grieve, R., Clancy, E., Klettke, B., van Dick, R., & Hernandez Bark, A. S. (2021). The Role of Individual Differences in Cyber Dating Abuse Perpetration. *CyberPsychology, Behavior & Social Networking*, 24(7), 457–463. https://doi.org/10.1089/cyber.2020.0687.

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

Meta (2021). Facebook Diversity Update: Increasing Representation in Our Workforce and Supporting Minority-Owned Businesses. https://about.fb.com/news/2021/07/facebook-diversity-report-2021/

Meta (2022). Embracing change through inclusion: Meta's 2022 diversity report. https://about.fb.com/wp-content/uploads/2022/07/Meta_Embracing-Change-Through-Inclusion_2022-Diversity-Report.pdf

Møller, K., & Robards, B. (2019). Walking Through, Going Along and Scrolling Back: Ephemeral mobilities in digital ethnography. *Nordicom Review*, 40(1), 95–109. https://doi.org/10.2478/nor-2019-0016

Nelson, R., Robards, B., Churchill, B., Vivienne, S., Byron, P. & Hanckel, B. (2022). Social media use among bisexuals and pansexuals: Connection, harassment and mental health. *Culture, Health & Sexuality*, https://www.tandfonline.com/doi/full/10.1080/13691058.2022.209221 3.

Nutall, L. (2020). Five technology design principles to combat domestic abuse. IBM, United Kingdom.

*Online Safety Act 2021* (Cth)

Reed, P., Bircek, N. I., Osborne, L. A., Viganò, C., & Truzoli, R. (2018). Visual Social Media Use Moderates the Relationship between Initial Problematic Internet Use and Later Narcissism. *The Open Psychology Journal*, 11(1). https://doi.org/10.2174/1874350101811010163

Robards, B. & Lincoln, S. (2020). *Growing up on Facebook.* Peter Lang.

Rosenfeld, M., Thomas, R., & Hausen, S. (2019). Disintermediating your friends: How online dating in the United States displaces other ways of meeting. *Proceedings Of The National Academy Of Sciences*, 116(36), 17753-17758. https://www.jstor.org/stable/26851209

Rowse, J., Bolt, C., & Gaya, S. (2020). Swipe right: The emergence of dating-app facilitated sexual assault. A descriptive retrospective audit of forensic examination caseload in an Australian metropolitan service. *Forensic Science, Medicine, and Pathology*, 16(1), 71–77. https://doi.org/10.1007/s12024-019-00201-7

Stardust, Z., Gillett, R., & Albury, K. (2022). Surveillance does not equal safety: Police, data and consent on dating apps. *Crime, Media, Culture.* https://doi.org/10.1177/17416590221111827

Statista. (2023). *Online Dating—Australia: Statista Market Forecast.* https://www.statista.com/outlook/dmo/eservices/dating-services/online-dating/australia

Suzor, N. (2019). Lawless: The secret rules that govern our digital lives. Cambridge University Press, United Kingdom.

The eSafety Commissioner (2022). *Australians' negative online experiences 2022.* https://www.esafety.gov.au/research/australians-negative-online-experiences-2022

Thierer, A. D. (2007). Social Networking and Age Verification: Many Hard Questions; No Easy Solutions. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.976936

Tinder Safety Centre. (2023). https://policies.tinder.com/safety/intl/en

Tinder Noonlight FAQs. (2023). https://www.help.tinder.com/hc/en-us/articles/360039260031-Noonlight-FAQs-

Van De Wiele, C., & Tong, S. T. (2014). Breaking boundaries: The uses & gratifications of Grindr. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (619-630). https://doi.org/10.1145/2632048.2636070

WESNET. (2020). Apple's new safety check - a tool for survivors. https://techsafety.org.au/blog/2022/06/22/apples-new-safety-check-a-tool-for-survivors/

WESNET. (2022). About the Technology Safety Summit. https://techsafety.org.au/techsummit/

WESNET and Tinder. (2023). Tinder Dating Safety Guide. WESNET. https://techsafety.org.au/wp-content/uploads/2023/02/Dating-Safety-Guide-single-page-for-printing.pdf

Wolbers, H., Boxall, H., Long, C., & Gunnoo, A. (2022). *Sexual harassment, aggression and violence victimisation among mobile dating app and website users in Australia.* Australian Institute of Criminology. https://doi.org/10.52922/rr78740

World Health Organisation. [WHO] (2016). Ethical and safety recommendations for intervention research on violence against women. WHO. https://apps.who.int/iris/bitstream/handle/10665/251759/9789241510189-eng.pdf