

Technology, gendered violence and Safety by Design



An industry guide for addressing technology-facilitated gender-based violence through Safety by Design

September 2024



Contents

01	Commissioner's foreword
03	Overview
05	What is technology-facilitated gender-based violence?
07	Who does it affect?
08	What is the impact?
09	Who are the perpetrators?
11	How digital products, services and data can be weaponised
23	Applying the Safety by Design principles
33	Conclusion
33	Want to share your Safety by Design approach?

Commissioner's foreword



Addressing gendered violence in the digital age

While I'm heartened by the growing swell of support for Safety by Design, I'm concerned it's not happening at the pace and speed so desperately needed.

I know from first-hand experience the tech sector is made up of some of the brightest and most innovative people motivated to harness tech for good.

But if we zoom out to what's happening at an industry level, it's clear too many design choices are failing to consider the safety and human rights of women, girls and diverse communities. It signals a lack of understanding of how technology is weaponised for gender-based violence.

While gender-based violence is a pervasive problem deeply rooted in many cultures, technology has made it easier to inflict harm at unfathomable speed and volume. It manifests in ways that reinforce harmful stereotypes, obliterate confidence, silence voices and undermine safety. This includes the objectification and sexualisation of women, threats of sexual violence and an obsession with family status. The prevalence of online vitriol is even higher if the person targeted lives with disability, is First Nations or culturally and linguistically diverse, or is gender diverse.

Technology can also be a tool of coercion, manipulation and control in situations of domestic and family violence. From innocuous child monitors to sophisticated tracking devices, abusers are using technology to inflict egregious harm on their partners and children.

And let's not forget how it can be used to demean anyone who has shared or posted their photos online. A recent example is the uptake of 'nudify' apps to create deepfakes. Using these apps costs a perpetrator virtually nothing, while the cost to the victim-survivor is lingering and incalculable.

The future is potentially even more chilling. The combination of full-sensory haptics, hyper-realistic augmented and virtual reality headsets, coupled with invasive and decentralised technologies could prove an insidious vector for sexual violence, causing profound harm and trauma.

This should give us all pause. And the clock is ticking.

That's why we need the immediate and wholesale adoption of Safety by Design across industry.


'Service provider responsibility', 'user empowerment', and 'transparency and accountability' are the key foundational pillars of Safety by Design, meaning the burden of safety should never fall upon the person targeted.

By applying a gendered lens to Safety by Design, companies would examine every feature and design aspect to minimise risks for women and girls at the design phase, engineering out potential misuse **before** launch. We believe the entire sector can and should take a much more assertive role in using their detection technologies to root out toxic and misogynistic conduct and content **before** they surface online. Powerful algorithms should not amplify gendered hate nor send people down harmful rabbit holes.

In this resource you'll find practical steps tech companies can take to prevent technology-facilitated gender-based violence through platform design. Also included are industry case studies, ranging from leveraging AI to detect harassment, to reducing unwanted interactions in private messages.

I hope you find this Safety by Design resource empowers you to create, design and build online products and services that exceed the safety needs and expectations of your clients and customers.

By acknowledging the gendered nature of online violence and prioritising Safety by Design, you're making platforms and services more inclusive and welcoming for everyone.



Overview

In 2018, the eSafety Commissioner (eSafety) launched Safety by Design, a critical initiative that outlines how and why technology companies should embed safety into technology products, from conception to launch and beyond.

eSafety did not work in isolation to produce the Safety by Design initiative. We conducted in-depth research and consulted with a broad spectrum of the technology ecosystem – industry, NGOs, advocates, parents and young people.

This collaboration yielded three foundational principles for online safety.

- 1. Service provider responsibility:** Tech companies must take responsibility for safety on their platforms and services – the burden should never fall solely on the user.
- 2. User empowerment and autonomy:** Users should have control over their online experiences and tools to protect themselves – these tools should be easy to find and use, and be effective.
- 3. Transparency and accountability:** Platforms should be open about how their products and services operate according to safety objectives and share innovations in safety.

These principles guided further discussion and research to develop robust risk assessment tools and resources. These tools have universal application – for providers of established online platforms and services to cutting-edge generative artificial intelligence (AI) applications.

This guide is a blueprint for technology companies; a way for them to understand and address [technology-facilitated gender-based violence](#) (TFGBV) so their services can be instruments for good, rather than weapons of hatred, misogyny and abuse. **It outlines how Safety by Design can help companies anticipate, detect and eliminate TFGBV before it happens.**

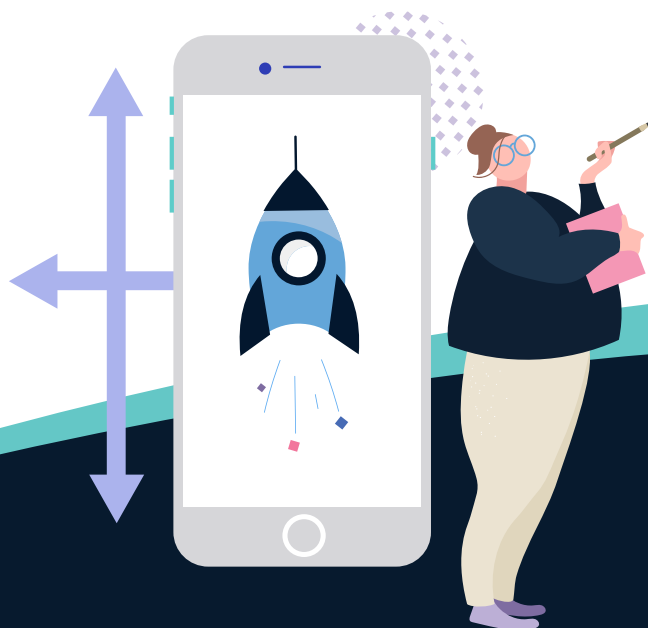


The guide includes case studies of technical interventions and platform policy responses developed with Bumble, Match Group, Meta, Telstra, TikTok, Twitch, and YouTube as well as Catherine Fitzpatrick, who worked closely with Australian banks to drive Safety by Design in banking. However, inclusion here does not imply a company has conquered TFGBV. **The fight against this abuse demands an ongoing and systemic commitment across the technology ecosystem.** By sharing best practices, we aim to inspire innovation and investment, ultimately creating a safer online world for everyone.

The prevalence of gender-based violence in our communities, amplified by the disturbing increase in technology-facilitated abuse, remains a pressing concern. The infamous ‘Gamergate’ – a year-long harassment campaign in 2014-15 on sites including 4chan and Reddit – serves as a chilling reminder of the digital abuse women and girls face, often escalating to sexualisation and even threats of rape and death.

What makes technology-facilitated gender-based violence distinct from other forms of gender-based violence is its scale, velocity and impact – using coordinated tactics known as pile-ons, volumetric attacks, dog-piling and brigading. It can be perpetrated remotely, cheaply, anonymously and through content that remains available indefinitely.¹

Due to the vast amount of data they collect, technology platforms are often the only entities that can identify the signals of these attacks before they happen or shut down systemically abusive accounts. Technology companies have an opportunity and responsibility to make sure their platforms are not exploited to facilitate this abuse.



What is technology-facilitated gender-based violence?

Gender-based violence is any form of physical or non-physical violence or abuse against a person or group because of biased or harmful beliefs about gender.²

TFGBV is where gender-based violence occurs online or through digital technology.

The impacts of TFGBV are far-reaching and often devastating, inflicting physical, sexual, psychological, social, political and economic harms, as well as other infringements of rights and freedoms.³ Technology is commonly used to facilitate gendered violence through:

Stalking and monitoring



Threats of violence, including rape and death threats



Coercive control



Trolling



Sexual harassment



Violent pornography



Image-based abuse / non-consensual intimate image sharing



Swatting



Online hate, including misogyny linked to violent extremism



Misinformation and disinformation



Doxing



Impersonation accounts⁴



²eSafety, [Gendered violence](#), Key Topics, eSafety website, 2023.

³UN Women and World Health Organisation Joint Programme on Violence Against Women, [Tech-facilitated violence against women: Towards a common definition](#), UN Women website, 2023 and World Economic Forum, Toolkit for Digital Safety Design Interventions and Innovations: [Typology of Online Harms](#)

⁴eSafety, [Gendered violence](#), eSafety website, 2023; The Global Partnership, [Technology-Facilitated Gender-Based Violence: Preliminary Landscape Analysis](#), Foreign, Commonwealth and Development Office, UK Government, 2023; S Bundtzen, ['Misogynistic Pathways to Radicalisation: Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence'](#), Institute for Strategic Dialogue, 14 September 2023.



International organisations are working towards an agreed definition of TFGBV. UN Women and the World Health Organisation convened an expert group in 2022 that proposed the following:

‘Technology-facilitated gender-based violence (TFGBV) is any act that is committed, assisted, aggravated, or amplified by the use of information communication technologies or other digital tools, that results in or is likely to result in physical, sexual, psychological, social, political, or economic harm, or other infringements of rights and freedoms.’

Who does it affect?

TFGBV primarily impacts women and girls but can also extend to LGBTIQ+ people. This is because of its roots in:

- gender discrimination – stemming from harmful attitudes, beliefs, stereotypes or behaviours about gender
- gender inequality – exploiting power imbalances in society where gender dictates how people are treated
- multiple forms of prejudice – including sexism, misogyny, homophobia, biphobia and transphobia.⁵

TFGBV can be experienced with other forms of discrimination, such as racism, ableism and ageism, which adds another layer or dimension to the abuse. This is sometimes called ‘intersectional discrimination’. For example, a First Nations woman’s experience of TFGBV may include sexism ‘intersecting’ with racism. In this way, **there is a higher risk of being the target of abuse if the woman is from a diverse background, is young or is living with a disability.** Women with high-profile jobs or leadership positions – such as politicians, journalists, activists, athletes – are also more likely to experience TFGBV.

Women and girls already experiencing family, domestic or sexual violence are especially susceptible to TFGBV. Additionally, adolescent girls and young women are more likely to experience TFGBV than other women.⁶



⁵eSafety, **Gendered violence**, Key Topics, eSafety website, 2023.

⁶eSafety, **Technology-facilitated abuse: family, domestic and sexual violence literature scan**, eSafety website, October 2023; UNFPA, **Making all spaces safe: technology-facilitated gender-based violence**, UNFPA website 2023; The Global Partnership, **Technology-Facilitated Gender-Based Violence: Preliminary Landscape Analysis**, Foreign, Commonwealth and Development Office, UK Government, 2023.

What is the impact?

TFGBV can inflict profound, long-lasting impacts on victim-survivors, affecting their physical, emotional and mental health. It can instil fear, paranoia and hypervigilance, and can severely damage personal and professional relationships and financial wellbeing.⁷

Our work with family, domestic and sexual violence services reveals a disturbing trend: violence pivots between online and offline environments. **TFGBV is often a precursor to physical violence.** A 2017 survey of 46 Australian women who survived technology-facilitated intimate partner stalking indicated they were also likely to endure other forms of violence within the same relationship. The study found that 82% of participants had suffered emotional abuse, 58% sexual abuse, 39% physical violence, and 37% financial abuse.⁸

In the gaming world, eSafety's research reveals that 11% of teens have seen or heard other players expressing or sharing misogynistic ideas. Participants said this resulted in some young women hiding their gender or leaving gaming environments overrun by misogyny.⁹

The harms extend far beyond the individual. TFGBV fuels discrimination, amplifying harmful norms around gender and sexuality, such as those that tolerate sexual violence. **It stifles diversity in public discourse¹⁰, and undermines fundamental human rights and freedom of expression.¹¹** For example, in a UNESCO survey, nearly three-quarters (73%) of female journalists said they had experienced online violence. Threats of physical (25%) and sexual violence (18%) were identified as common forms of gendered abuse. Fearing for their safety, 30% self-censored on social media and 20% withdrew from all online interaction. Others reported missing work to recover (11%), quitting their jobs (4%) and abandoning journalism altogether (2%).¹²

⁷eSafety, [Technology-facilitated abuse: family, domestic and sexual violence literature scan](#), eSafety website, October 2023.

⁸eSafety, [Technology-facilitated abuse: family, domestic and sexual violence literature scan](#), eSafety website, October 2023.

⁹eSafety, [Levelling up to stay safe: Young people's experiences navigating the joys and risks of online gaming](#), eSafety website, February 2024.

¹⁰F Stevens et al., [Women are less comfortable expressing opinions online than men and report heightened fears for safety: Surveying gender differences in experiences of online harms](#), arXiv, 27 March 2024.

¹¹The Global Partnership, [Technology-Facilitated Gender-Based Violence: Preliminary Landscape Analysis](#), Foreign, Commonwealth and Development Office UK, 2023.

¹²UNESCO, [The Chilling: Global trends in online violence against women journalists: Research discussion paper](#), UNESCO, p12, April 2021.

Who are the perpetrators?

Understanding who the perpetrators are and their motives is critical to any assessment of TFGBV. The United Nations Population Fund (UNFPA) developed a valuable tool – a map of threat actors that shows the diverse sources of online threats (figure 1).

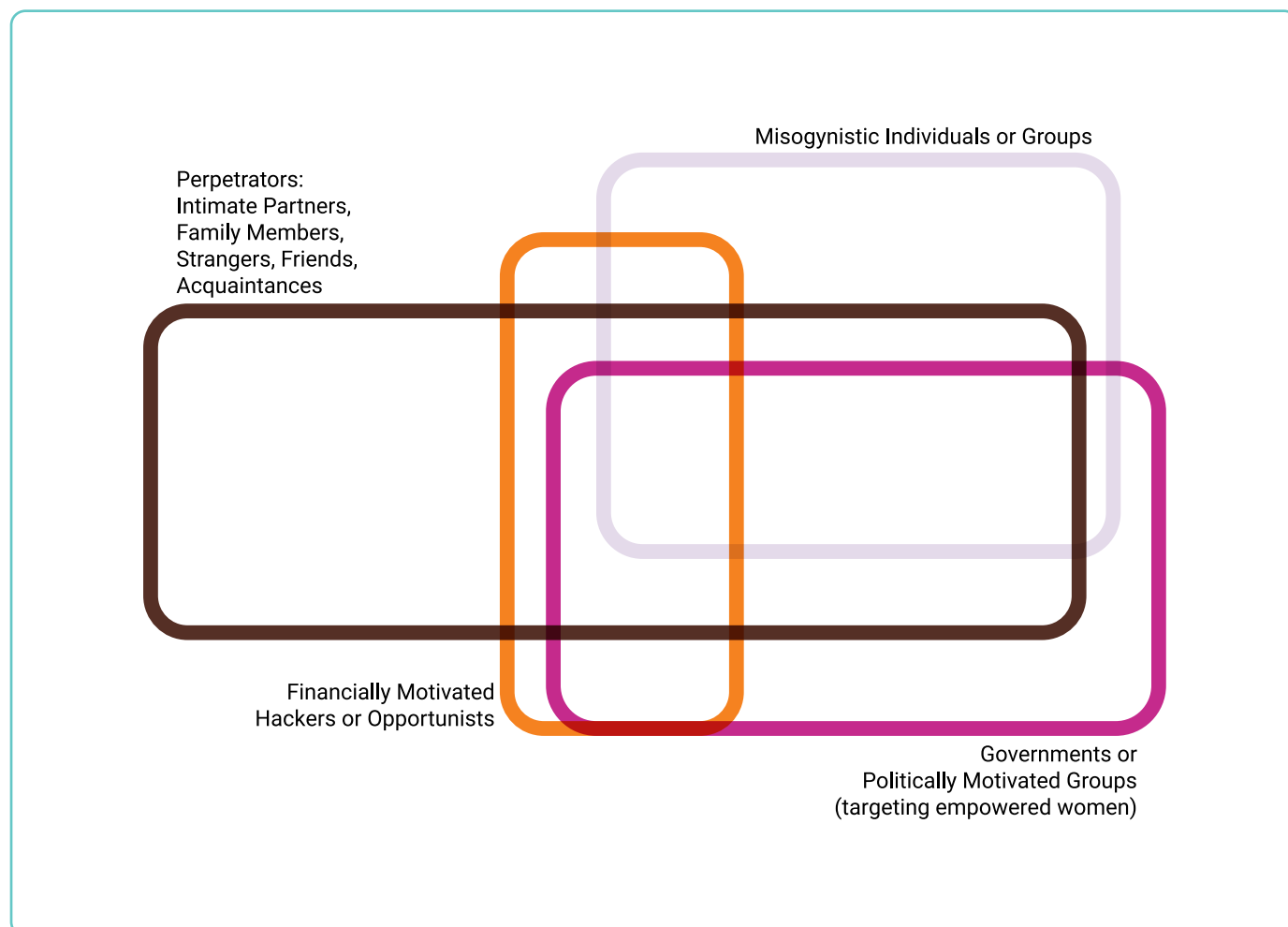


Figure 1: Threat actor mapping from UNFPA Guide on the Safe and Ethical Use of Technology to Address Gender-Based Violence and Harmful Practices: implementation summary

¹³UNESCO, **The Chilling: Global trends in online violence against women journalists: Research discussion paper**, UNESCO, p12, April 2021.

¹⁴MD Schwartz, **'Masculinities, Sport, and Violence Against Women: The Contribution of Male Peer Support Theory'**, Violence Against Women, 2021, 27(5), 688-707, doi.org/10.1177/1077801220958493.

¹⁵C Jones, V Trott and S Wright, **'Sluts and soyboys: MGTOW and the production of misogynistic online harassment'**, New Media & Society, 2020, 22(10), 1903-1921, doi:10.1177/1461444819887141; S Zimmerman, **'The Ideology of Incels: Misogyny and Victimhood as Justification for Political Violence'**, Terrorism And Political Violence, 2022, doi:10.1080/09546553.2022.2129014.

Some forms of TFGBV are perpetrated by people unknown to the victim-survivor. These acts often originate from anonymous accounts and can be conducted through organised or semi-organised networks of abusers who target high-profile women.¹³

These online networks, while diverse in their specifics, often fall under the umbrella of the 'Manosphere' – a term coined to represent a spectrum of online communities which propagate sexist, misogynistic and violence-supporting attitudes.¹⁴ These communities, often bound by a sense of entitlement over women's bodies and a desire to prevent their autonomy, include Involuntary Celibates (Incels), Men's Right's Activists, Pickup Artists, Men Going Their Own Way, Father's Right's activists, and even some 'geek/nerd' groups.¹⁵

TFGBV can also manifest in the context of family and domestic violence, particularly as a tool of [coercive control](#). In such scenarios, digital technologies often help the abuser exert and maintain power over the other person, eroding their confidence, freedom and independence. A literature scan by eSafety, reveals that **in Australia one in three experiences of technology-facilitated abuse are perpetrated by a current or former intimate partner**, with women more likely to experience the abuse than men.¹⁶

The motivations behind TFGBV are complex and varied. Financial gain, political agendas and extremist ideology can all fuel online abuse. Other common drivers include anger, a desire for control, an attitude of sexual entitlement, and even a twisted sense of entertainment for the perpetrator.¹⁷



¹⁶eSafety, [Technology-facilitated abuse: family, domestic and sexual violence literature scan](#), eSafety website, October 2023.

¹⁷Australia's National Research Organisation for Women's Safety (ANROWS), [Technology-facilitated abuse: National survey of Australian adults' experiences](#), ANROWS website, 2022; International Centre for Research for Women (ICRW), [Technology-facilitated gender-based violence: What is it, and how do we measure it?](#), ICRW website, 2018.

How digital products, services and data can be weaponised

Perpetrators often use multiple tactics and behaviours across different digital platforms and services to target victim-survivors. No product is immune to being weaponised to facilitate TFGBV, but these are some of the most common:

- **Social platforms:** Social media, gaming platforms, dating apps and messaging services all become potential battlegrounds.
- **Online accounts:** Banking, email and other online accounts can be hijacked to exert control and inflict financial hardship.
- **Websites and image boards:** These include 'revenge porn' websites that encourage users to upload nude or sexual images of others.
- **Tracking devices, applications and spyware:** Perpetrators use these as tools to monitor, stalk and control other people.
- **Smart, internet-connected devices (Internet of Things):** Abusers can hack devices such as smart speakers, home camera/security systems and doorbells to intimidate others.
- **Generative AI applications:** This includes open-source models that perpetrators can use to create deepfakes or other forms of digital harassment.¹⁸
- **Augmented and virtual reality hardware and software:** These technologies introduce new avenues of potential abuse.

Social platforms

Using social platforms can have positive and negative consequences. While fostering connection and communication, they also have features and tools abusers can weaponise to perpetrate TFGBV. The following features are especially susceptible to abuse.

- **User-generated content:** Perpetrators can weaponise posts, livestreams, AI, chatbots, avatars and more to harass, intimidate and humiliate victims. Often in a coordinated manner.
- **Communications tools:** Comments, direct messages, groups chats and voice chats can become channels for abuse. Perpetrators often use encrypted services, as well as fake and anonymous accounts to evade detection and being held to account.
- **Recommender systems:** Designed to enhance online experiences by suggesting relevant content, these systems can inadvertently or intentionally amplify harmful material.
- **Reporting mechanisms:** Mechanisms to report and mitigate inappropriate behaviour can be ineffective, slow to respond, or not actioned, creating a culture of impunity for abusers.

¹⁸Graphika, [A Revealing Picture](#), Graphika website, 2023.

📁 Case studies: preventing unsolicited nudes via direct messages

Sending nude or sexual content to someone without their consent is a form of sexual harassment, sometimes known as 'cyberflashing'. This can occur via in-app messages, texts, email, or device features such as AirDrop and Nearby Share.

While sending unsolicited nudes is commonly associated with online dating, it happens in other contexts as well, including interactions between strangers.

Findings of a UK survey of women and marginalised genders published in June 2023 showed 34% of respondents reported experiencing cyberflashing, but this was higher with younger age groups. Of all respondents aged 16-24, almost half (47%) reported being a victim of cyberflashing, followed by 45% of respondents aged 25-34.¹⁹

Technical intervention by Bumble

Bumble developed Private Detector™, an AI detection tool that automatically blurs potential nude images shared in chat, while still giving the user the option to view the image if they choose to do so.²⁰ The person who receives the blurred image is also prompted with an option to report the sender, making it easy for users to flag unsolicited photos. This feedback loop helps to improve moderation policies and processes.

Developing the Private Detector tool was a cross-functional effort:

- The research team worked to understand the issues associated with unsolicited photos and their impact on the customers' experience.
- Developers worked alongside the product team to build a flexible and customised tool.
- The Trust and Safety team received additional training on consent and context to improve the response to customer reports of sexual harassment.

Bumble has made Private Detector [open source](#) to encourage other tech companies to adapt and create features that combat online abuse and harassment.

Technical intervention by Meta

Meta, in collaboration with experts from its Safety Advisory Council, introduced a message gating system to offer users protection from unwanted interactions in Instagram DM and Messenger.²¹

This feature limits the number of messages and types of content a non-follower can send to another user. The sender can only send one text message with no images, voice notes or calls. Additional messages and content can only be sent if the recipient accepts the message request, preventing them from receiving repeated unwanted contact and images.

The feature operationalises the concept of consent because the recipient is in control of who can communicate with them. Meta initially tested the feature during the 2023 FIFA Women's World Cup and later rolled it out permanently due to its success in disrupting negative behaviours and its popularity among users.

¹⁹Communia UK, [Report on Women's and Marginalized Genders' Social Media Experiences](#), Communia website, 16 December 2022.

²⁰L Matney, ['Bumble's 'Private Detector' uses AI to detect dick pics'](#), Techcrunch, 25 April 2019.

²¹D Milma, ['Instagram tests feature to block explicit images in direct message requests'](#), The Guardian, 28 June 2023; Meta, ['Giving Teens and Parents More Ways to Manage Their Time on Our Apps'](#), Meta, 27 June 2023.

Recommender systems

Recommender systems, a common feature on social platforms, can amplify harmful content and conduct. They can boost discriminatory content that fuels bias, sexism and misogyny. This can harm those exposed to such content and those who are the subject of the content, and – on a broader societal level – **serve to normalise prejudice, hate and violence**.²²

One example is the case of coordinated hate and misogyny directed against actor Amber Heard during the Depp v Heard defamation trial. Not only did algorithms contribute to circulating violent hate and harassment, but algorithms also acted to provide ‘cultural capital’ within misogynistic and far-right groups online – further mobilising others to engage in TFGBV against Heard.²³ Such examples of algorithmic misogyny and TFGBV not only attack individual women and their stories, but they also spread and normalise harmful narratives, including rape myths. These narratives often deter women from speaking out about their abuse and limit their equal participation in online spaces.²⁴

Case study: search interventions for gendered hate

Technical intervention by TikTok

One of the ways TikTok aims to mitigate TFGBV is by reducing the discoverability of search results in areas that are violative of its Community Guidelines. For example, users searching for content that TikTok deems violative of these guidelines in relation to hate – such as content relating to specific individuals who promote hateful ideologies – will not be shown directly related results and related terms will not appear on suggested search lists or predictive text.

This intervention supports TikTok’s scaled removal of content by limiting the discoverability of hateful content. TikTok publicly reported that in the period July–September 2023 it removed 92% of Hate Speech and Hateful Behaviour Content proactively, before anyone reported it, and 87% of this content was removed within 24 hours. For Violent and Hateful Organisations and Individuals, TikTok says it removed 98% of content proactively, before anyone reported it, and 83% was removed within 24 hours.

²²eSafety, [Recommender systems and algorithms – position statement](#), eSafety website, December 2022.

²³S Robinson and E Hiltz, ‘[Platformed misogyny in Depp v Heard: #justiceforjohnny and networked defamation](#)’, *Feminist Media Studies*, 2024, doi:10.1080/14680777.2023.2284107.

²⁴E Rennie, ‘[What a lying slut: the \(re\)production of rape myths in online misogyny towards women disclosing their experiences of rape through the #MeToo movement](#)’, *Journal of Gender-based Violence*, 2022, 7(2), doi.org:10.1332/239868021X16699044856526.

Reporting tools

Reporting tools, while designed to flag harmful content and behaviour, can be weaponised as part of coordinated attacks against victim-survivors. This misuse, known as mass reporting, is a technique of coordinated abuse deployed as a form of brigading.²⁵ Brigading involves a group of people working together to harass or silence others through high-volume techniques such as mass commenting, posting or reporting. Sometimes facilitated by the mass creation of fake and imposter accounts, brigading has become a key feature of ‘networked gaslighting’ of women journalists.²⁶

These high-volume attacks, also known as ‘volumetric’ or ‘pile on’ attacks, can occur across platforms.²⁷ **It is crucial platforms act on signals detected by their own systems to stem these attacks.** As much of this abuse is cross-platform, knowledge-sharing with other service providers is also critical. This becomes particularly important in the face of increasingly widespread access to generative AI, and the potential to exploit AI models and their outputs to automate harassment at scale.



²⁵P CS Andrews, ‘[Social Media Futures: What Is Brigading?](#)’, Tony Blair Institute for Global Change, 10 March 2021.

²⁶UNESCO, [The Chilling: Global trends in online violence against women journalists: research discussion paper](#), UNESCO, p12, April 2021.

²⁷eSafety, [Basic Online Safety Expectations Reasonable Steps](#), eSafety website, 2024.

Tracking applications

Even technology designed to enhance safety for women and girls can be misused against them. For example, there are applications that allow a person to share their location with trusted family or friends as a safety measure. In the hands of abusers, these apps become tools of control, stripping victim-survivors of their right to privacy. eSafety research reveals that in 2022, 16% of Australians had someone electronically track their location without their consent.²⁸

This misuse of technology often surfaces in family and domestic violence situations. The NSW Crime Commission found **one in four people who purchased tracking devices in NSW in 2023 have a history of domestic violence**. The Commission also found many customers purchased a device in the days after a protection order had been enforced.²⁹ Also, an Australian survey on the use of technology to facilitate family and domestic violence revealed that, according to practitioners, women were being tracked with GPS via apps or devices ‘all the time’ (28%) and ‘often’ (38%).³⁰ This highlights the urgent need for robust measures to prevent such misuse.

Smart devices

Our homes are becoming increasingly smart, with the average Australian household boasting about 24 internet-connected devices, such as smart home assistants, security systems, fitness trackers, item trackers and children’s watches.³¹

For people experiencing family and domestic violence, these devices often become tools of abuse. Many have specific design features that inadvertently create vulnerabilities.³² For example, they give perpetrators access to information about the victim-survivor’s movements, which they can use to coerce and control them.

However, [new research](#) suggests a variety of low-cost modifications can significantly reduce this risk. These include adding visual and noise indicators when a device or feature is activated and having access data logs that are easy for users to find and check.

Australians are rapidly adopting AI-powered technology and home robotics, with the average number of smart devices per home expected to reach 33 by 2027.³³ This growth in interconnected devices creates an increasingly complex landscape for managing the risk of technology misuse.

²⁸eSafety, [Australians’ negative online experiences 2022](#), eSafety website, 2022.

²⁹NSW Crime Commission, [Project Hakea: Criminal use of tracking and other surveillance devices in NSW \[PDF\]](#), NSW Crime Commission, 25 June 2024.

³⁰WESNET, [The Second National Survey on Technology Abuse and Domestic Violence in Australia](#), WESNET website, 2020.

³¹Telsyte, [Australia’s smart home market set to crack \\$2.5 billion, driven by AI, energy savings and security](#), Telsyte, website, 20 March 2024.

³²A Brown, D Harkin and LM Tanczer, ‘[Safeguarding the “Internet of Things” for Victim-Survivors of Domestic and Family Violence: Anticipating Exploitative Use and Encouraging Safety-by-Design](#)’, Violence Against Women, 2 January 2024, doi:10.1177/10778012231222486.

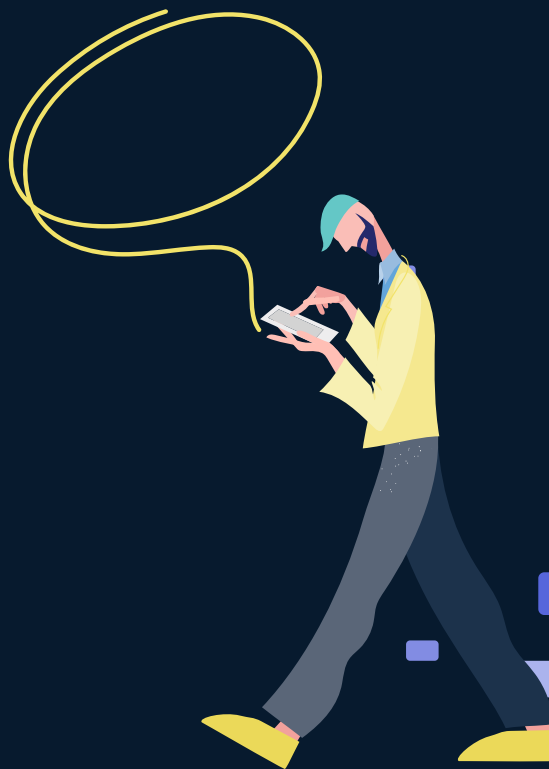
³³Telsyte, [Australia’s smart home market set to crack \\$2.5 billion, driven by AI, energy savings and security](#), Telsyte website, 20 March 2024.

📁 Case study: warning users of stalking and tracking via smart devices

Technical intervention by Apple and Google

Location tracking devices help users find personal items via Bluetooth wireless signals. These popular devices can be – and have been – abused for the unwanted tracking of individuals. Regulators like eSafety wrote to relevant companies expressing concerns about the potential weaponisation of the devices upon their initial release, in 2021. Ideally, a Safety by Design approach would have been taken in advance of their design, development and deployment.

Over time and in response to these harms being realised, Apple and Google collaborated on an industry-first specification to combat unwanted tracking across iOS and Android. The specification, made up of a set of best practices and protocols for manufacturers whose products have built-in location-tracking capabilities, aims to establish a standard for applications that can detect and warn users about suspected unwanted tracking. Input from various safety and advocacy groups, such as The National Network to End Domestic Violence, has been integrated into the development of the specification.³⁴



³⁴Apple, 'Apple and Google lead initiative for an industry specification to address unwanted tracking', Apple Website, 2024.

Generative AI

There are benefits and risks with generative AI. On the positive side, online service providers have used AI to detect and remove harmful content, increasing the accuracy of moderation³⁵ and large language models (LLMs) such as Open AI's GPT-4. These have the potential to streamline the creation and customisation of platform content policies.³⁶

However, generative AI also presents risks and harms. At the individual level, it can generate and amplify harmful gender-based content, such as 'deepfake pornography' for image-based abuse. On a societal level, it can contribute to the generation and amplification of content that promotes gender-based bias and discrimination.

Without strong safeguards, generative AI can be used to commit or amplify TFGBV. This can include:

- generating non-consensual intimate images (NCII), such as deepfake pornography and 'nudifying' images
- producing personalised gendered hate speech, bullying, abuse and other forms of harassment and manipulation at scale³⁷
- creating synthetic personal histories with false and harmful gender-based narratives that defame or denigrate an individual³⁸
- normalising the absence of women and diverse voices, values and perspectives while amplifying and reinforcing gender stereotypes and biases.

Deepfake image-based abuse

Deepfake image-based abuse, including deepfake pornography, represents a significant and increasing threat, with estimates suggesting it comprises **96% of all deepfake videos online**.³⁹ Anyone who has had their image digitally captured can become a target for the creation of fake non-consensual intimate images. Several studies report that about **99% of this content depicts women**.⁴⁰

The ease and accessibility of creating deepfakes is concerning. Powerful AI apps, including 'nudify' apps, are freely available and weaponised by perpetrators, with devastating impacts on victim-survivors. Additionally, open-source code repositories can provide a blueprint for bad actors, with the source code and algorithms underpinning deepfake programs and apps that are readily available.

³⁵Cambridge Consultants, '[Use of AI in Online Content Moderation](#)', Ofcom website, 2019.

³⁶L Weng, V Goel and A Vallone, '[Using GPT-4 for content moderation](#)', Open AI blog, 15 August 2023.

³⁷eSafety, '[Generative AI – position statement](#)', eSafety website, 2023.

³⁸R Chowdry and L Dhanya, '[Your opinion doesn't matter anyway": Exposing technology-facilitated gender-based violence in an era of generative AI](#)', UNESCO, 2023.

³⁹H Ajder, G Patrini, F Cavalli and L Cullen, '[The State of Deepfakes: Landscape, Threats, and Impact \[PDF\]](#)', Deeptrace, September 2019.

⁴⁰Home Security Heroes, '[2023 State of Deepfakes: Realities, Threats and Impact](#)', Home Security Heroes website, n.d.



Data bias and harmful outputs

Generative AI models are trained on data and this process carries risks related to how that data is collected, stored and shared. Studies reveal that AI image generators exhibit gender biases, overwhelmingly producing images of professionals, politicians and leaders as male.

For example, one study found that in a sample of 99 images from nine popular text-to-image generators, professionals (doctors, lawyers, engineers, scientists) were represented as men in 76% of the images and as women in only 8%. Doctors had the smallest representation of women (7%), despite women comprising nearly half of all doctors in the latest Organisation for Economic and Co-operation and Development country findings.⁴¹ These data-based gender biases perpetuate the notion that men are natural leaders and decision makers, but women are not.

AI applications such as ChatGPT present a complex situation. Perpetrators can misuse these tools to learn tactics on new and emerging forms of technology-facilitated abuse, as well as how to hide their crimes from law enforcement.⁴² However, they also hold promise for victim-survivor support. With proper training on trauma-informed and victim-survivor-centred data, ChatGPT and similar applications could offer victim-survivors vital information on staying safe.⁴³

eSafety's [position statement on Generative AI](#) offers valuable guidance for the technology industry. It outlines specific Safety by Design interventions and emerging best practices. It also provides meaningful, actionable and achievable industry guidance to minimise existing and emerging generative AI harms, supporting user safety and empowerment.

⁴¹A Gorska and D Jemielniak, '[The invisible women: uncovering gender bias in AI-generated images of professionals](#)', *Feminist Media Studies*, 2023, doi:10.1080/14680777.2023.2263659.

⁴²R Chowdry and L Dhanya, '[Your opinion doesn't matter anyway](#)': [Exposing technology-facilitated gender-based violence in an era of generative AI](#), UNESCO, 2023.

⁴³P Novitzky, J Janssen, B Kokkeler, '[A systematic review of ethical challenges and opportunities of addressing domestic violence with AI-technologies and online tools](#)', June 2023, <https://doi.org/10.1016/j.heliyon.2023.e17140>.

Case studies: identifying synthetic media

Generative AI tools can be weaponised to facilitate online impersonation and the production of gendered disinformation. Women in politics often find themselves at the forefront of targeted attacks designed to objectify them and reinforce gender stereotypes.⁴⁴

Policy response by YouTube

YouTube has expanded its privacy policy request process to allow users to request the removal of AI-generated or other synthetic or altered content that looks like them. When evaluating these requests YouTube will consider whether:

- the content is parody or satire, or has other public interest value
- the person making the request can be uniquely identified
- the content is realistic
- the content features a public official or well-known individual engaging in a sensitive behaviour, such as criminal activity, violence or endorsing a product or political satire.

YouTube now also requires creators to disclose to viewers when they've created altered or synthetic content that appears realistic (meaning it could easily be mistaken for a real person, place, scene or event), including through the use of AI tools. Creators who consistently choose not to disclose this information may incur penalties like removal of the content or suspension from the YouTube Partner Program.

Content will still be removed from YouTube if it violates its Community Guidelines, which apply to all forms of content on the platform including synthetic and altered content. For example, YouTube's nudity and sexual content policy does not allow pornography, regardless of how it is generated.

Technical response by YouTube

When a creator discloses their content to be meaningfully altered or synthetically generated and it looks realistic it triggers:

- an explanatory label in the description panel to indicate that sounds or visuals were synthetically edited or digitally generated
- an explanatory label directly on the video in the player window for content about sensitive topics such as health, news, elections or finance.

In some cases, YouTube may add a label even where the creator hasn't made a disclosure, especially when the content discusses sensitive topics such as health, news, elections or finance.

Explanatory labels will also be placed on all content created by any of YouTube's Generative AI products and features.

⁴⁴A Blatnik, [An Overlooked Threat To Democracy? Gendered Disinformation About Female Politicians](#), Women in International Security (WIIS) website, n.d.

Technical intervention by the Content Authenticity Initiative and Coalition for Content Provenance and Authenticity

The Content Authenticity Initiative (CAI) is an Adobe-led cross-sector effort to address misinformation and content authenticity issues at scale. It is a community of hundreds of technology companies, NGOs, academics, creators, technologists, journalists and activists. The CAI collaborates on developing content attribution standards and open-source tools around content authenticity.

The Coalition for Content Provenance and Authenticity (C2PA) complements the work of the CAI by providing the end-to-end open technical standards for creators, editors, publishers, media platforms and consumers. These standards provide the ability to trace the origin of different types of media. C2PA specifications are supported by companies including Microsoft, Meta, Adobe, Open AI and Leica, and have the ultimate goal of allowing content credentials to be applied from the moment an image is captured (for example on a Leica camera) or otherwise created (for example through Open AI's tools).

Although the labelling of content through C2PA is opt-in (and image generation tools used to create deepfake pornography, for example, can't be compelled to comply with it), and provenance data may be stripped or swapped by bad actors, these collaborative efforts spanning the technology ecosystem and involving other key stakeholders represent an important step in addressing manipulated content risks.



Virtual, augmented and mixed reality

Virtual and immersive spaces can reflect and exacerbate harms that occur in real-world settings, such as unwanted interactions or simulated physical assault. [Research by eSafety](#) found 9% of Australian metaverse users had experienced unwanted haptic touch.⁴⁵ **These virtual worlds also present unique dangers, where users can create or encounter offensive virtual objects, avatars or signage.** Alarming, 6% of users reported that someone had created a sexually explicit avatar or image of them for non-consensual interactions.⁴⁶

A survey of more than 600 users of HTC Vive, Oculus Rift, PlayStation Virtual Reality and Microsoft Windows Mixed Reality found **49% of women respondents described experiencing sexual harassment in immersive spaces.** This includes groping, stalking, catcalling, being shown a lewd photograph, or hearing sexually explicit comments. Notably, 36% of male respondents reported similar experiences.⁴⁷ Female avatars or those with feminine voices are more susceptible to harassment.⁴⁸

Interactions in the metaverse are designed to feel real and can feel personal and private. However, the fleeting nature of these interactions, happening in real-time, means it can be difficult to document and report bad behaviour.⁴⁹ As such, **deploying safety features by default in the metaverse may be more effective at reducing harm than moderating behaviours as they occur or afterwards.**⁵⁰



⁴⁵Further reading: L Blackwell, N Ellison, N Elliott-Deflo and R Schwartz, '[Harassment in Social Virtual Reality: Challenges for Platform Governance](#)', PACM on Human-Computer Interaction, 2019, 3 (100), doi:10.1145/3359202.

⁴⁶eSafety, '[The metaverse: A snapshot of experiences in virtual reality](#)', eSafety website, 2023.

⁴⁷J Outlaw, '[Virtual Harassment: The Social Experience of 600+ Regular Virtual Reality \(VR\) Users](#)', Virtual Reality Pop, 4 April 2018.

⁴⁸L Blackwell, N Ellison, N Elliott-Deflo and R Schwartz, '[Harassment in Social Virtual Reality: Challenges for Platform Governance](#)', PACM on Human-Computer Interaction, 2019, 3(100), doi:10.1145/3359202.

⁴⁹M M Ashraf, '[Gender-based Abuse on the Metaverse: The New Internet is Being Coded on a Toxic Palimpsest](#)', Bot Populi, 23 January 2023.

⁵⁰M M Ashraf, '[Gender-based Abuse on the Metaverse: The New Internet is Being Coded on a Toxic Palimpsest](#)', Bot Populi, 23 January 2023.

Case study: addressing harassment in the metaverse

Technical interventions by Meta for Horizon Worlds

To give people more control over their experience in Horizon Worlds, Meta introduced a feature called Personal Boundary. This gives users an invisible boundary that creates more space between people so others can't come too close and makes it easier to avoid unwanted interactions. In the testing phase of a virtual reality game in Horizon Worlds this feature was made available for users to activate when they felt threatened. During this time a woman tester [reported she was groped](#) while playing, prompting Meta to make the safety features of Horizon Worlds easier to find and use. Meta made several tweaks based on community feedback and gave users more customised controls on who was allowed within a user's Personal Boundary. The default setting for Personal Boundary is now turned on for people a user doesn't know, and off for friends. Users have the option to then turn it on or off for everyone.

Meta has also introduced a tool called Pause that immediately takes users to a space away from other people and their surroundings. From here people can mute, block or report content or people.

Applying the Safety by Design principles

The principles of Safety by Design, agreed to through extensive research and consultation, are:

1. **Service provider responsibility**
2. **User empowerment and autonomy**
3. **Transparency and accountability**

The principles provide practical guidance for platforms and services as they incorporate, assess and enhance user safety. They are supported by realistic, actionable and achievable measures enabling companies to implement Safety by Design regardless of their size or maturity. At the heart of these principles is a human-centric approach that prioritises the safety and rights of users, while also considering their needs and expectations. This approach positions user safety as the third pillar in the development process for all online and digital technologies, alongside privacy and security.

Service provider responsibility

The burden of safety should never fall solely on the user, especially those who are victim-survivors of gender-based violence. Tech companies must take proactive steps to understand, assess and address online harms in the design and delivery of their platforms and services.

This means anticipating potential risks during online interactions and designing features to eliminate potential misuse, reducing people's exposure to harms.

Steps to take:

1. **Assign responsibility:** Appoint individuals or teams and make them accountable for creating, evaluating, implementing and operating women's safety policies. Make sure these individuals or teams understand TFGBV and how it may manifest on the platform or service, so it is considered throughout the policy lifecycle.
2. **Develop guidelines:** Consider forms of TFGBV in the development of community guidelines, terms of service and moderation procedures and make sure these are implemented fairly and consistently. The [banking](#) and [Twitch](#) case studies are examples of how to consider forms of TFGBV in community guidelines and terms of service.
3. **Establish reporting mechanisms:** Include forms of TFGBV within reporting mechanisms that are readily accessible for users to flag concerns and report violations at the point they occur. These should be part of an infrastructure that supports internal and external triaging, clear escalation pathways, and reporting on all user safety concerns. The [Bumble](#) case study is an example of how to make reporting options available at the point a violation occurs and establish supporting infrastructure to handle reports.

- 4. Set clear protocols:** Ensure clear internal protocols for reports of TFGBV are in place for engaging with regulators, law enforcement (where there is immediate risk of harm), and support services (where victim-survivors need support and guidance on next steps).
- 5. Implement detection processes:** Establish processes to quickly detect, surface and action TFGBV content, contact and conduct with the aim of preventing or reducing harm. This may involve a combination of automated and AI solutions, human moderation, and tools from safety technology providers. Processes may include:
 - deploying detection technologies such as photo matching, machine learning and AI to proactively detect the different forms of TFGBV at scale, and with high rates of accuracy
 - removing, demoting or sending detected forms of TFGBV content to a human team for further review, in line with community guidelines and terms of service
 - ensuring predictive text in search does not suggest words, phrases or figures associated with TFGBV – the [TikTok](#) case study is an example of detecting and limiting specific search terms, with the aim of reducing harm.
- 6. Conduct risk management:** Undertake continuous risk management and impact assessments to identify and assess the risk of different forms of TFGBV for each product or service and key components such as recommender systems. Ensure these are documented and kept up to date.
- 7. Implement social contracts:** Do this at the point of registration, outlining the safety duties and responsibilities of the service, user and third parties. This should explicitly include the obligation not to engage in gender-based violence.
- 8. Consider security, privacy and safety:** Integrate security by design, privacy by design and safety by design throughout product development. This includes considering the ongoing confidentiality, integrity, and availability of personal data. TFGBV-related data is particularly sensitive to misuse and abuse, which requires platforms and services to secure it appropriately and understand what data is collected, how it's stored, for how long and who has access to it.



Case studies: disrupting non-consensual intimate image sharing

Technical intervention by Twitch

Twitch has intentionally omitted the ability for users to send images in private and in-stream chats to reduce the risk of NCII, or other unwanted sexual imagery, being shared through these channels.⁵¹

Policy and educational measures by Twitch

Despite existing measures to prevent the sharing of NCII, in January 2023 a Twitch streamer broadcasted that they were viewing synthetic non-consensual exploitative images (or synthetic NCEI, commonly referred to as deepfake pornography) of women streamers.⁵² This incident prompted Twitch to review the gaps in its processes through which the NCEI content was live streamed and how women streamers could be supported. Twitch sought feedback from streamers and consulted with subject matter experts, including the Cyber Civil Rights Initiative (CCRI) and the UK Revenge Porn Helpline.

Following this review, Twitch:

- updated its Adult Sexual Violence and Exploitation policy to make it clear that intentionally promoting, creating or sharing synthetic NCEI is prohibited and can result in an indefinite suspension on the first offence
- clarified that their Adult Nudity policy includes synthetic NCEI, which means that even if NCEI is shown only briefly or, for example, shown to express outrage or disapproval of the content, it will be removed and result in enforcement action
- hosted a 'Creator Camp' program to educate Twitch streamers about NCEI material and what steps they should take if they encounter it online. The program was led by the UK Revenge Porn Helpline and informed by the abovementioned consultations.⁵³

StopNCII.org empowers victims of abuse to take back control

[Stop NCII.org](https://stopncii.org) is a free tool designed to support victims of NCII abuse. It works by generating hashes from intimate images or videos selected by an individual on their device, and then sharing those hashes with participating companies so they can detect matches within their systems and remove them. Hashes are created on the individual's device and only the hash, not the associated video or image, is sent to StopNCII.org.

⁵¹Twitch consultation with eSafety, 22 November 2023.

⁵²Twitch, **Addressing Explicit Deepfake Content**, Safety News, Twitch website, 8 March 2023.

⁵³Twitch, **Addressing Explicit Deepfake Content**, Safety News, Twitch website, 8 March 2023.

User empowerment and autonomy

The dignity of users is paramount. Tech products and services should prioritise user wellbeing and uphold fundamental consumer and human rights. This is critical in the context of TFGBV, where abusers often seek to undermine the dignity and autonomy of victim-survivors.

Abuse can be intersectional, impacting users in multiple ways for various reasons. Technology can exacerbate societal inequalities. To counter this, platforms and services need to have meaningful conversations with marginalised groups at higher risk of abuse. This ensures their features and functions are accessible to all.

Steps to take:

- 1. Implement safety measures:** Provide technical measures and tools that allow users to manage their own safety. These should be set to the most secure privacy and safety levels by default. These are some examples:
 - **Restricting who can contact a user and how** – the [Meta](#) case study is an example of giving users technical tools to manage unwanted interactions.
 - **Providing robust transparency around device access, functions and data**, particularly for devices with cameras, microphones and location tracking that can accept multiple users. In cases of family and domestic violence, this can allow a victim-survivor to be alerted that a former partner still has remote access to a device.⁵⁴ [Safeguarding the ‘Internet of Things’ for Victim-Survivors of Domestic and Family Violence: Anticipating Exploitative Use and Encouraging Safety-by-Design](#) explores the application of abusability testing and safety by design approaches to Internet of Things devices.
 - **Allowing creators to control who can comment on their content** – the [TikTok](#) case study is an example of comment control options being made available to creators.
- 2. Establish clear protocols:** Define clear protocols and consequences for service violations, including those related to TFGBV. These consequences should serve as effective deterrents. The [banking](#) case study is an example of establishing meaningful deterrents for service violations.
- 3. Prompt safer interactions:** Use technical features to prompt users towards safer interactions and mitigate risks of TFGBV at key points within the service. The [Tinder](#) case study is an example of deploying technical features to prompt safer interactions.
- 4. Provide support functions:** Build in support functions and constructive feedback loops for users reporting TFGBV. These systems should clearly and promptly inform users about the status of their reports, including outcomes and avenues for appeal.
- 5. Design reporting forms:** Design reporting forms that allow users to report multiple pieces of content simultaneously and provide context about their situation and the harmful content or conduct being reported. This context may include details of harassment across multiple platforms or services, combined with offline threats and abuse.
- 6. Evaluate design features:** Throughout the risk management and impact assessment lifecycle, evaluate all design and function features to ensure risks for all users, especially those vulnerable to TFGBV, are mitigated before public release.

⁵⁴A Brown, D Harkin and L.M. Tanczer, ‘[Safeguarding the “Internet of Things” for Victim-Survivors of Domestic and Family Violence: Anticipating Exploitative Use and Encouraging Safety-by-Design](#)’, Violence Against Women, 2024, doi:10.1177/10778012231222486.

Case studies: tackling gender-based harassment and bullying

Technical intervention by TikTok

To empower users, TikTok provides optional comment controls for those who wish to minimise their exposure to negative comments and manage incidents. For example, users can use comment controls to filter offensive comments, specific keywords and spam, even if these are not violative of TikTok's Community Guidelines.

Users can also use the 'Filter All Comments' tool. When enabled, comments aren't displayed unless the video's creator approves them. Leveraging these controls enables users to minimise their exposure to specific commentary and limits the ability of other users to perpetrate harms such as gendered hate and doxing (for example by blocking comments that share their home address or mobile number publicly) through the platform's comment section.

TikTok prompts and encourages users to engage with these tools and to regularly review their account settings.

Educational measure by TikTok

TikTok also has a [Bullying Prevention Guide](#) which helps users to identify the various types of behaviours that individuals or groups use to intimidate, degrade or disempower others and to understand how to implement controls that might help prevent such behaviours. The guide also offers tips to manage the impact of bullying.

Third-party blocking tool used on Twitter

Block Party was a third-party tool developed by Tracy Chou to help individuals escape targeted harassment campaigns on then-Twitter by letting users automate the process of blocking bad actors and creating block lists. For example, if a particular tweet was breeding harassment or abuse, users could automatically block anyone who liked or retweeted that post. Users could also review content they had chosen to block or mute at a later time of their choosing or elect a trusted friend to do it for them.

Through a mix of customisation to suit individual user needs and ready-to-use default settings, Block Party empowered people to take control of their experience on Twitter and promoted more inclusive conversations. Changes to Twitter's API pricing in 2023 led to Block Party being placed on an indefinite hiatus. But Ms Chou has pivoted with the introduction of [Privacy Party](#), which is in Beta for multiple platforms.

📁 Case study: detecting sexual harassment and abusive language

Technical intervention by Tinder

Tinder's 'Does This Bother You?' feature uses AI to detect potentially offensive messages received by Tinder users. If the system detects such a message, the recipient receives an in-app prompt asking them 'Does this bother you?'. If the recipient indicates 'Yes', the app directs them to the reporting form. Tinder says that reports of inappropriate messages increased by 46% for members who had seen the prompt.⁵⁵ The feature is informed by what users have previously reported and continuously improves as the system collects more data.

Tinder introduced another similar feature in 2021, the 'Are You Sure?' prompt. This feature uses key words to detect harmful language in draft messages, such as overtly sexual or violent language. Once harmful language is detected, the sender receives an in-app prompt asking them if they are sure about sending the message, prompting them to pause and consider its contents.⁵⁶ More key words, phrases, and emojis that may be considered harmful have been added to the feature over time.⁵⁷ Since the initial rollout more than 500 million messages have either been edited or deleted by users after the feature was triggered.⁵⁸



⁵⁵Tinder, [Tinder Introduces Are You Sure?, an Industry-First Feature That is Stopping Harassment Before It Starts](#), Tinder Pressroom website, 2021.

⁵⁶Tinder, [Tinder Introduces Are You Sure?, an Industry-First Feature That is Stopping Harassment Before It Starts](#), Tinder Pressroom website, 2021.

⁵⁷A Malik, '[Tinder rolls out new safety features, including an Incognito Mode](#)', Techcrunch, 7 February 2023.

⁵⁸Tinder correspondence with eSafety, 7 February 2024.

Transparency and accountability

Transparency and accountability are hallmarks of a robust approach to safety. They assure users that platforms and services are following their stated safety goals. They also help to educate and empower users about how to address their safety concerns. Transparency and accountability can help victim-survivors know their rights and protections and incentivise industry to improve standards in response to TFGBV.

Publishing data on how companies enforce their policies, and the effectiveness of safety features or innovations, allows for clear assessment of what works. If interventions improve user safety or deter TFGBV, these innovations should be shared and adopted more widely.

Steps to take:

- 1. Integrate TFGBV into user safety considerations:** Incorporate the risk of TFGBV into user safety considerations, training and practices that are integral to the roles, functions and working practices of all individuals associated with the product or service. The [banking](#) and [Telstra](#) case studies are examples of this approach.
- 2. Maintain accessible user safety policies:** Ensure that user safety policies, terms and conditions, community guidelines and processes about user safety – including those relating to TFGBV – are accessible by users of all abilities, easy to find, regularly updated and easy to understand. Send users periodic reminders about these policies and tell them of major changes or updates through targeted in-service communications. The [Twitch](#) case study is an example of educating creators about policy changes.
- 3. Work with experts and advocates:** Collaborate openly with experts, victim-survivor advocates and key stakeholders in TFGBV to develop, interpret and apply safety standards that are effective and appropriate.
- 4. Publish transparent reports:** Publish annual assessments on the prevalence of different forms of TFGBV and the effectiveness of measures in place to address them. These reports should include meaningful analysis of relevant metrics, such as incidence data, user reports, moderation efforts, and compliance with community guidelines and terms of service. In published reports, be mindful to:
 - avoid details that perpetrators could exploit to circumvent safety measures
 - omit details of specific incidents that could identify an individual or group, including location, sexual orientation, religion or other sensitive information that could put them at risk.
- 5. Innovate and collaborate:** Continuously invest in technologies that mitigate TFGBV and share best practices, processes and tools with others. The [banking](#) case study is an example of sharing tools to address TFGBV.

Case study: blocking abusive messages in banking

In 2019, an investigation by Commonwealth Bank Australia (CBA) identified that over a 3-month period, more than 8000 customers had received multiple low-value deposits, often less than \$1, with potentially **abusive messages in the transaction descriptions**. While some of these transactions appeared to be jokes between friends, many included serious profanity, threats, and clear references to domestic and family violence.⁵⁹

CBA shared these findings with the Australian Banking Association, a trade association comprised of 20 Australian banks, and has since introduced measures to prevent and respond to financial abuse, including abuse which occurs through the weaponisation of digital banking products as a form of TFGBV.⁶⁰ These measures include but are not limited to the following:

- **updating its acceptable use policy** to reflect that customers found to be using their online banking apps to engage in unlawful, defamatory, harassing or threatening conduct, or promoting or encouraging physical or mental harm or violence against any person, may have their transactions refused or access to digital banking services suspended or discontinued⁶¹
- **enabling user-reporting** of abusive transaction descriptions⁶²
- implementing **automatic detection, filtering and blocking** of abusive transactions, such as those with abusive payment descriptions⁶³
- using AI and machine learning technology alongside the automated actions to enhance abuse detection - the system learns and considers context and existing relationships between customers when assessing whether to block a transaction identifying, for example, differences between joke transactions between friends and small value transactions from an abusive partner who may be using the payment description as a method of unwanted communication⁶⁴
- making its **machine learning techniques available for free** to any bank in the world.⁶⁵

CBA blocks approximately 400,000 transactions annually through its transaction description filtering system, and its AI and machine learning system detects over 1,500 cases per year. These outcomes are a practical example of how embedding safety in the design of technology can prevent harm on a mass scale.

⁵⁹P Smith, '**CBA targets abusive messages through online banking**', Australian Financial Review, 4 June 2020.

⁶⁰C Fitzpatrick, **Designed to Disrupt: Reimagining banking products to improve financial safety**, The Centre for Women's Economic Safety (CWES), Discussion Paper 1, 2022.

⁶¹Commonwealth Bank Australia (CBA), **CBA to launch police referral pilot in NSW to address tech-facilitated abuse**, Newsroom, CBA website, 2023.

⁶²Commonwealth Bank Australia (CBA), **CommBank Next Chapter**, CBA website, n.d.

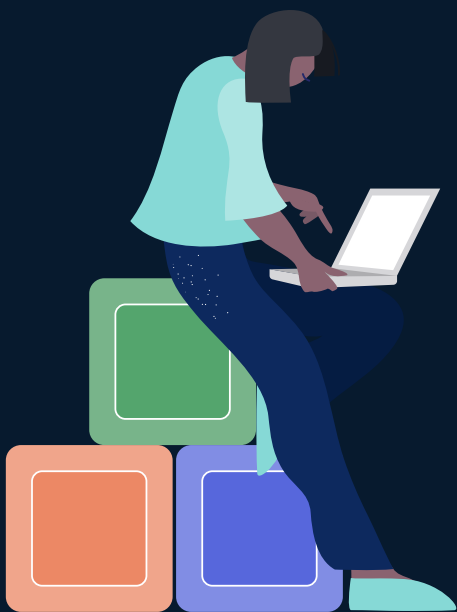
⁶³Commonwealth Bank Australia (CBA), **CBA to launch police referral pilot in NSW to address tech-facilitated abuse**, CBA website, 2023.

⁶⁴P Smith, '**CBA turns to AI to root out abusive payments**', Australian Financial Review, 11 October 2021.

⁶⁵Commonwealth Bank Australia (CBA), **In a world first, CBA shares its artificial intelligence model to help reduce technology-facilitated abuse**, CBA website, 2023.

All of Australia's major banks have now implemented similar technology systems and policy measures to proactively detect, filter and respond to abuse facilitated through transaction payment descriptions and other products, including the following:

- ANZ developed an algorithm to prevent its digital payment services being used to send abusive messages.⁶⁶
- NAB has over 1300 words and phrases blocked from being used in the NAB app.⁶⁷
- Westpac monitors outgoing payments, blocking certain transactions containing inappropriate or offensive language in real-time. The sending customer is notified and required to amend the language in order to proceed with payment. Westpac customers can also report abusive messages received via transactions.⁶⁸ Other banks are introducing similar systems.⁶⁹
- 19 Australian banks have updated their terms and conditions with a zero-tolerance policy for financial abuse and unacceptable account conduct.⁷⁰ This sets expectations of acceptable customer behaviour and challenges the condoning of violence against women.⁷¹ It is also a key recommendation in [Design to Disrupt: Reimagining banking products to improve financial safety](#), published by the Centre for Women's Economic Safety
- several banks, including CBA and Westpac, are also implementing Safety by Design training for staff working in design areas.⁷²



⁶⁶M Dalling, [Using design to combat economic abuse](#), bluenotes ANZ website, 18 October 2023.

⁶⁷Digital Nation, [‘NAB increases measures to block abusive transaction descriptions’](#), Digital Nation, 12 April 2022.

⁶⁸Westpac Bank Australia, [Westpac strengthens safeguards against abusive messages sent via payment transactions](#), Westpac website, 2021.

⁶⁹C Fitzpatrick, Flequity Ventures consultation with eSafety, 21 November 2023.

⁷⁰Commonwealth Bank Australia (CBA), [Electronic banking terms and conditions \[PDF\]](#), CBA, 6 December 2023; NAB, [Changes to Personal Transaction and Savings Products \(Addendum\) Terms and Conditions, Effective 01 November 2023 \[PDF\]](#), NAB 2003; Westpac Bank Australia (Westpac), [New measures to combat financial abuse](#), Westpac website, 26 July 2023.

⁷¹OurWatch, [‘Change the Story: A shared framework for the primary prevention of violence against women in Australia \(second edition\)’](#), OurWatch, 2021

⁷²C Fitzpatrick, Flequity Ventures consultation with eSafety, 21 November 2023; Westpac consultation with eSafety, 24 January 2024.

📁 Case study: embedding TFGBV considerations in training and working practices

Employee training by Telstra

[The Second National Survey on Technology Abuse and Domestic Violence](#) (2020) showed that being bombarded with texts, emails and messages or receiving threats are the most common forms of technology-facilitated abuse. Checking women's phones without permission and using text, email and instant messaging to check where women are, is also common.

To support victim-survivors, Telstra has a trained Safe team.⁷³ The Safe team:

- is a group of specialists trained to support customers affected by domestic and family violence, and provides a direct point of contact through voice and messaging
- will check how a customer wants to proceed and will not ask for proof of the situation before providing assistance
- supports customers to make informed decisions about their accounts and services and to safely make changes to any existing account settings and services including changes of ownership and/or activation of new services.

Customers access the Safe team through a dedicated phone number available on the Telstra website or are transferred from other agents via a voice or messaging channel if they identify as a victim-survivor and require specialist assistance. Customers can also request that the Safe team contact them on a number and at a time that is safe to do so.

For added safety, on Telstra's domestic and family violence support page, there is a quick exit button that redirects to Google. This page is also not captured in browser history.

⁷³Telstra, [Domestic and family violence assistance](#), Telstra website, n.d.

Conclusion

By applying a gendered lens to the Safety by Design principles, we show how tech companies can play a more proactive role in combating TFGBV.

Through in-depth research, and in consultation with Industry, eSafety has surfaced good technical and policy interventions that demonstrate how implementing Safety by Design can lead to safer online environments for women, girls and diverse communities. Though ultimately, safer digital spaces benefit society as a whole.

We encourage all tech companies to use this guide to comprehensively take a fresh look at their policies, practices and products, to create a safer and more inclusive online space for everyone.

Want to share your Safety by Design approach?

Safety by Design is a journey and we are always looking for new ideas, perspectives and best online safety practices.

Collaborating is an important way we can continue to reflect the diverse needs of the technology ecosystem and make the online world safer for everyone.

Our collaborations span a range of industries and specialisations, and include working with:

- business leaders
- industry safety specialists
- tertiary educators and students
- government organisations and NGOs
- community organisations
- anyone looking to engage with our expertise.

For more information about collaborating on our Safety by Design initiatives, please contact us at industry@esafety.gov.au.



